Predicting Future Serious Criminal Behavior in a Cohort of Juveniles

February 22, 2002

Abstract

In the mid 1980s, the Office of Juvenile Justice and Delinquency Prevention (OJJDP) established Habitual Offender Units (HOUs) in 13 U.S. cities, including Philadelphia. By assigning those juveniles with the highest risk for future serious criminal behavior to more experienced prosecutors, the OJJDP hoped to avert future crime. However, no city's HOU based its selection criteria upon research on what factors actually predict becoming a serious offender. The current study used data from police records for a cohort of Black and white juveniles born in 1958 and living in Philadelphia. By examining the relationship between aspects of current and past offenses and future crimes, it was hoped that more effective selection criteria for the HOU could be determined. This report summarizes predictor variables for the cohort, describes how the relevant outcome variables were derived, and finally presents and critiques models to predict which juveniles are most at risk for high levels of criminal activity and severity of offenses as adults, based on the first and second contacts they have with the police.

In the mid 1980s, the Office of Juvenile Justice and Delinquency Prevention (OJJDP) established Habitual Offender Units (HOUs) in 13 U.S. cities, including Philadelphia. By assigning those juveniles with the highest risk for future serious offending to more experienced prosecutors, the OJJDP hoped to avert future crime. The selection criteria used to determine which youths would be given priority prosecution was different in each of the 13 cities, but no city based its criteria upon research on what factors actually predicted serious future offending. Generally, juveniles who accumulated a prespecified number of specific serious charges against themselves were assigned to the HOU, but it was unknown how accurate these criteria were in correctly identifying the juveniles most likely to become serious offenders as adults. Without knowing whether the program was targeting the right juveniles, it was impossible to determine whether its limited resources were being put to the most effective use.

1.1 Overview and Objectives

Data on juvenile and adult crimes were collected for a group of 1436 Black and white males born in 1958 who resided in Philadelphia from at least age 10 to 18 and who had at least one contact with the police. Because the crimes committed by this group were already documented, it was possible to see how well specific aspects of a juvenile's criminal behavior predicted his future crimes. In addition, this cohort's police contacts all occurred before the HOU was instituted in Philadelphia; it was possible to examine their criminal histories without intervention, whereas using a more current sample would have meant trying to control for the effect of this intervention.

1.2 Study Design and Data Collection

For each juvenile in the cohort, police contacts were identified using the "rap sheet" maintained by the Juvenile Aid Division of the Philadelphia Police Department. The rap sheet was used to locate the police investigation report that provided even more detailed information about that contact. Some additional information was taken from the police arrest report. The cohort's juvenile contacts spanned the years from 1968 to 1976. Information about adult arrests in Philadelphia was obtained from the computer files maintained by the Court of Common Pleas. The digital equivalent of the rap sheets, these computer files were used to locate the more detailed police investigation and arrest reports, up to the year 1985. Philadelphia Family Court records were used to obtain information on case dispositions. Race and ethnicity information was obtained from school records. The youth's home address was used to broadly categorize his SES.

I obtained this data from the National Archive of Criminal Justice Data, maintained by the Inter-University Consortium for Political and Social Research. The final dataset I used for building my models was based on a combination of the original variables as well as various transformations of these variables. This is described in detail in Sections 1.7.4 of the Appendix.

An important point to mention about the data in its current form is that each case represents a police contact, not an individual's entire police record.

1.3 Data description and Univariate Analyses

1.3.1 Predictor Variables

Current Offense: Seriousness, Charges, etc.

A series of dummy variables indicated whether the current crime fell into each of nine categories, which are in order of decreasing severity: (9) serious assaults such as homicide and aggravated assault, (8) nonserious assaults, (7) robbery, (6) sexual assault and other sex crimes, (5) weapons violations, (4) drug offenses such as possession, use, and sale, (3) serious property crimes such as larceny/theft, burglary, and motor vehicle theft, (2) other property crimes such as receiving stolen property, forgery, counterfeiting, fraud, and embezzlement, and (1) all other crimes. These categories correspond broadly to the six-category crime-seriousness scale developed by Gottfredson, Warner, and Taylor (1988). The largest number of cases in the dataset (1792 cases) were categorized as involving serious property crimes, other property crimes, and other crimes. Next, with 592 cases, were drug offenses. It was most common for a particular case to be put into three distinct crime categories.

I created a new variable to give a value to the maximum seriousness of the crime categories for each case. This discrete variable ranged from 1 to 9, in order of increasing seriousness (corresponding to the numbers above). For example, if a case was categorized as involving both robbery and weapons violations, it would receive a score of 7. Most cases in the dataset received a 3, for serious property crimes, although there were a large number of cases in more serious categories as well (see Section 1.7.2 of the Appendix).

The dataset also contained a scale variable calculate by study researchers to indicate overall seriousness of each case. This was a continuous variable which ranged in this dataset from 1 to 87.3, with the data centered around 9. Because parallel boxplots of this variable split by the derived maximum seriousness variable above did not show a clear relationship between the two, I included both in further analyses.

Over all police contacts, about 63% resulted in two to four charges being brought against the juvenile. It was less common to receive only one charge or more than five charges. Almost 90% of police contacts resulted in arrest.

Victims of Current Offense

Out of all the contacts in which the number of crime victims was known, 2746 had no victims, 2973 had one, and 318 had two or more.

I created another scale variable based on the level of victimization indicated by five dummy variables. The scores and corresponding interpretation of this variable is as follows: (0) no or mutual victimization, in which there is no harm is done to anyone without their consent, (1) tertiary victimization, which refers to any diffuse victimization extending to the community at large, such as offenses against the public order, (2) secondary victimization, in which the victim is impersonal, such as a large commercial establishment, (3) primary victimization, in which there is an individual victim but no face-to-face interaction, such as residential burglary or auto theft, and (4) face-to-face victimization, such as rape, assault, and robbery. The police contacts in this dataset generally involved a high degree of victimization; 2334 cases involved primary victimization, and 1978 involved face-to-face victimization.

There was a variable in the dataset which simply indicated whether there was any injury at all involved with the current offense; in about one fifth of cases there was. Another set of variables indicated the level of injury inflicted upon the victim, and I again combined these into one scale variable in which higher scores indicate higher levels of injury, defined as follows: (0) no injury or no victimization, (1) minor harm, (2) treated or discharged, (3) death or hospitalization. In general, the level of injury was low or nonexistent. Only 198 cases involved death or hospitalization.

The majority of cases did not involve physical intimidation or intimidation with a weapon, although they may have involved verbal intimidation.

The mean age of known victims was 32.11, and ages ranged from 1 to 97 years. There were over twice as many male victims as females in the dataset, and victims were about equally likely to be white as nonwhite.

Weapons Used in Current Offense

The categories of weapons indicated in the dataset were knives, firearms, and other weapons. The majority of crimes involved no weapon or an unknown weapon. A greater number of cases involved knives (580) over firearms (402). Very few cases had more than one type of weapon.

Offenders in Current Offense

Most crimes in the dataset were committed by more than one offender. Although all juveniles in the dataset are male, sometimes they had help from female offenders. However, this occurred in less than 2% of the known cases. In about 10% of known cases, the juvenile had help from someone over 18 years old.

Setting for Current Offense

Most crimes occurred in public places, although relatively large numbers of crimes also occurred in business and residential locations. It was about equally likely for crimes to be committed inside or outside.

Police Response to Current Offense

The reason for police response to a given event was most often that they had received some type of complaint, although often the police had actually observed the offense. When police action was in response to a complaint, the most typical complainant was a stranger to the offender; it was least likely for the complainant to be a friend or family member.

Prior Events

Prior offenses were categorized using the same nine categories that were used for current offenses, but in this case the categories were for *all* prior events, not necessarily just one crime. The breakdown for categories was similar to that for current crimes, however. I also created a "maximum seriousness" variable analogous to the one for the current crime. Again, among those who had committed at least one prior offense, it was most common for prior offenses to be placed in three of the statutory categories. Again, seriousness scales by the study designers were included for the overall seriousness of the first prior event and the overall seriousness of the most recent prior event.

There was a large amount of missing data in the variables indicating the legal disposition of prior events. In my analysis I let an unknown disposition be the baseline category and had indicator variables for cases that had been adjudicated and involved confinement, adjudicated without confinement, and not adjudicated. Of the known cases, the majority were not adjudicated.

The most common prior crimes were nuisance crimes and status violations.

Personal Characteristics of Offender

Most contacts were with juveniles in the middle range of socio-economic status. Few were in the top 15%. The dataset had information for Black and white juveniles only. 5124 contacts were with Black juveniles, and 1178 were with white juveniles. The median age of offenders ranged from about 15.5 years at the first contact to 16.5 years at the eighth contact. At all contacts there were many low outliers in terms of age, with a minimum value of five and a half years. The minimum for a juvenile's age at his first prior offense was slightly over two years old. The median for this variable was slightly over thirteen years.

1.3.2 Outcome Variables

Original Variables in the Dataset

There were 80 variables in the dataset designated as outcome variables. Forty of these were continuous, and 40 were dichotomized versions of the continuous variables. Among the 40 continuous variables, 20 of them dealt with crimes committed from the time of the current contact to the individual's 18th birthday, and 20 of them dealt with crimes committed from the time of the current contact to the individual's 27th birthday. I only used this latter set of more inclusive variables.

Now, among this set of 20 variables, there were four categories determined by types of crime: all criminal events, UCR index crimes, UCR violent index crimes, and UCR index crimes plus weapons, drugs, and other sex crimes. For each of these crime categories, there were five outcome variables. These were:

1. Rate per year of police contacts

2. Average seriousness per police contact

3. Average seriousness per year of police contacts

4. Average number of criminal charges per police contact

5. Average number of criminal charges per year

Data Reduction: Principal Components Analysis

There was a considerable amount of collinearity in the outcome variables; the R^2 values for regressing each of the 20 outcomes on all the others ranged from 0.874 to 0.999. It seemed likely multiple outcome variables were measuring similar underlying dimensions, since they could be grouped by type of crime and whether the variable was measured per year or per police contact. In order to determine the structure of the relationships between outcome variables and potentially reduce the number of relevant outcomes for analysis, I conducted a principal components analysis on the 20 variables (see section 1.7.3 of the Appendix). The first two components had clear interpretations and accounted for over 80% of the variance. All the crime categories tended to "hang together" in terms of loadings on the first two components. Specifically, the first component had high loadings for the rate per year of police contacts, the average seriousness per year of police contacts, and the average number of criminal charges per year, where these measures were for each of the four crime categories. In contrast, the second factor had high loadings for the average seriousness per police contact and the average number of criminal charges per police contact.

I interpreted these components to mean that the first is measuring is something related to a yearly measure of criminal activity, whereas the second is measuring something related to the severity of crime for each police contact. In light of this, I created two new variables from the component scores. For both of these new variables, higher scores indicate greater levels of crime. Noting that the factor loadings suggested that the principal component scores were approximately an average of the related variables, I created two new variables corresponding to those means. Plotting these two mean variables against their respective principal components scores (see Figure 1.14, in the Appendix), the means for the "activity" variables were a good approximation to the first principal component sores, but the means for severity were not as good an approximation to the second principal component scores. Therefore, I decided to use the principal component scores themselves for the analysis. However, one should bear in mind that the interpretation of the outcome variables is similar to simply taking the mean of the related variables.

If having more police contacts is predictive of future criminal behavior, then we would expect the distributions of these outcomes to change across police contacts. This is in fact the case; see Figures 1.1 and 1.2.

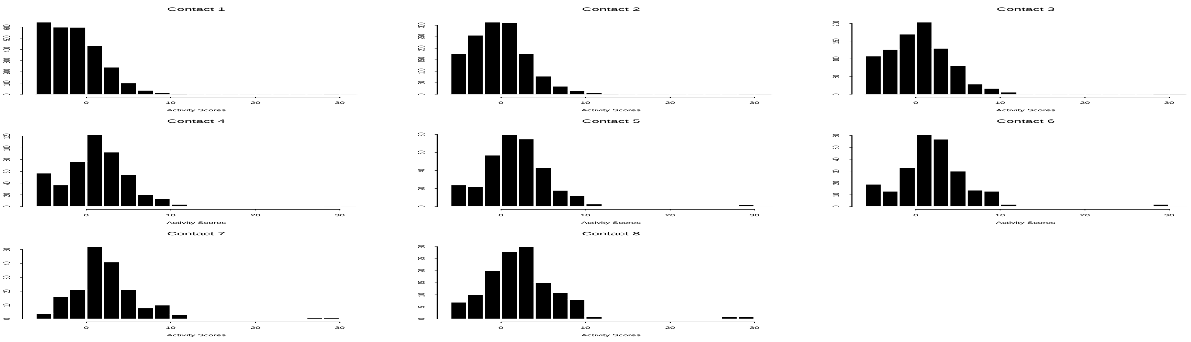

Figure 1.1: Distributions of "Activity" Principal Component Scores Across Police Contacts

Finally, because the research objective was to be able to predict the worst future offenders, I dichotomized the two outcome variables, activity and severity, by splitting the data at the 75th percentile. Thus, the operational definition of a serious offender was someone who fell into the top 25% of all offenders. This corresponded to the assertion by the study designers that the percentages assigned by the HOU prosecutors to identify the "high risk" group converged on the top quartile.

Actually, saying that a "1" for these risk indicator variables means that an individual falls into the top 25% of criminal activity or criminal severity over all individuals in the study is not quite precise, since each case represents a police contact, not an individual (many individuals have repeat contacts). However, doing a separate principal components analysis and dichotomization at each contact would not be appropriate either, because then the meaning of the outcome variables would be that an individual was in the top 25% of offenders, but only among those juveniles who had a greater or equal number of police contacts.

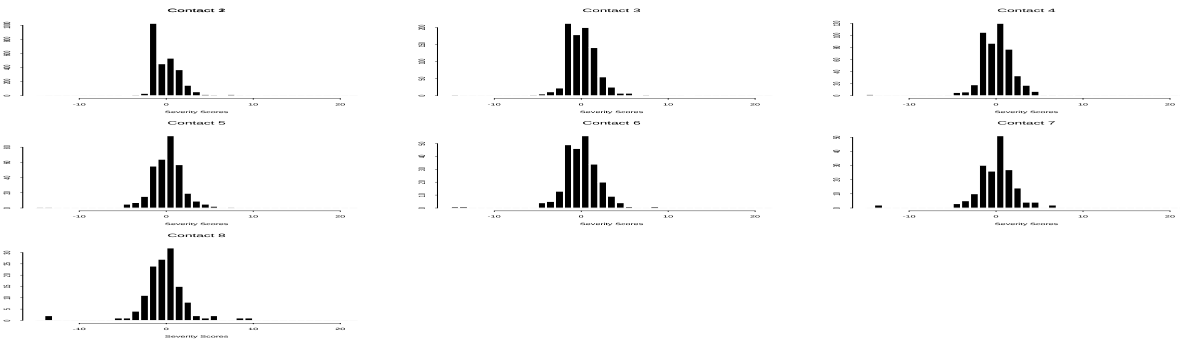

Figure 1.2: Distributions of "Severity" Principal Component Scores Across Police Contacts

1.4 Methods and Results: Model Building

Below I summarize and interpret logistic regression models to predict scoring in the "high risk" group for criminal activity and severity of crimes based on predictor variables taken from the first and second police contacts. The method I used to generate each model was similar. Because I had so many predictor variables, I relied on a stepwise procedure to eliminate potential models from consideration. I used AIC to rank models because it has been shown that it tends to include more variables than the true model actually contains. Then, I iteratively removed variables which were deemed insignificant by the likelihood ratio test when these variables are added last to the model. Please see the Appendix, Sections 1.7.5 and 1.7.6, for more detail on the models selected by the stepwise procedure.

To build these models I used 75% of the original data. That is, for both the first and second contact, a randomly selected 25% of cases were set aside for purposes of model validation, and the rest of the data was used for model selection.

Note that each model predicts the log odds of being a "high risk" juvenile:

$$\hat{L} = \log\left(\frac{\hat{P}(high)}{1 - \hat{P}(high)}\right)$$

where $\hat{P}(high)$ indicates the predicted probability of falling into the top quartile in terms of activity or severity of future crime. Predicted probabilities can be found by reversing this equation:

$$\hat{P} = \frac{1}{1 + e^{-\hat{L}}}$$

1.4.1 Models for First Contact

Activity

The model for criminal activity based on the first police contact is as follows:

$$\begin{aligned}\hat{L} \quad = \quad & -0.976 + 0.523(CURARRST) + 0.490(PRIEVNT1) + 1.013(PRIEVNT2) + \\ & 0.516(PRIVANDL) + 1.011(OFDRRACE) - 0.163(OFNDRAGE)\end{aligned}$$

Thus, the model indicates that being arrested for the current offense, having committed prior offenses, having vandalized, and being Black all significantly increase the log odds of falling into the top quartile of future criminal activity. In contrast, the older one is at the first police contact, the lower the log odds of falling into this high-risk group.

The conditional effects plot in Figure 1.3 shows the effect of age on the predicted probability of being in the high-risk group, for those with different numbers of prior events. All other variables in the model are held constant. Note the negative relationship between age and predicted probability, and the higher probabilities for those with greater numbers of prior criminal events.

Severity

The model for severity of crime based on the first police contact is as follows:

$$\begin{aligned}\hat{L} \quad = \quad & -1.958 + 0.118(CURVICSC) + 0.608(OFDRRACE) + 0.183(PRIEVNT1) + \\ & 0.382(PRIEVNT2) - 0.166(CURCH2.4) + 0.251(CURCH5..) - \\ & 0.259(CURINSID) + 0.046(CUROUTSI)\end{aligned}$$

Higher scores on the victimization scale predict higher log odds, as does being Black and having prior events. The baseline number of charges is one in this model, so the odds ratio for those with two to four charges is $e^{-0.166} = 0.847$ compared to those with one charge; their odds of being in the top quartile of crime severity are lower. In contrast, the odds ratio for those with five or more charges is $e^{0.251} = 1.286$ compared to those with one charge; their odds are higher. So it seems that having an average number of charges brought against you (two to four charges was the average for this group of juveniles) is predictive of less criminal behavior, compared to having either one charge or more than five charges. Finally, the baseline group in terms of location is for the location

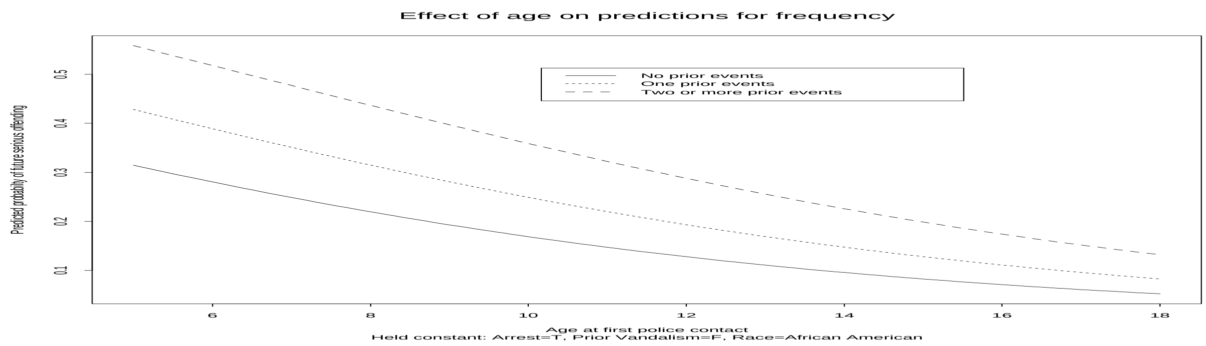

Figure 1.3: Conditional Effect of Age on Predicted Probability of Being High Risk for Activity

to be unknown. Knowing that the crime is either outside or inside predicts higher log odds, although being inside has a greater effect. It is difficult to say how the group of juveniles whose crimes were known to have been committed either outside or inside might be different from those with an unknown location. One might speculate that having an unknown location means police were less careful to record details about the crime, which would perhaps be the case for less serious crimes. However, if this were the case we might also expect the variables relating to seriousness to have been significant in the model.

1.4.2 Models for Second Contact

Activity

The model for criminal activity based on the second police contact is as follows:

$$\begin{aligned}\hat{L} \;=\; & 1.460 + 0.812(CURARRST) + 0.229(FIREARM) + 0.711(KNIFE) + \\ & 0.095(OTHWEAP) + 0.548(PRIEVNT2) + 1.612(ADJUDCONF) + \\ & 0.191(ADJUDNOT) + 0.237(NOTADJUD) + 0.693(OFDRRACE) - \\ & 0.312(OFNDRAGE) - 5.758(CURFEMOF)\end{aligned}$$

The variables that predict a higher log odds of being in the top quartile of criminal activity are being arrested for the current crime, carrying any kind of weapon (knives have the greatest effect), having more than 1 prior event, having a prior legal record (being both adjudicated and confined has the greatest effect), and being Black. In contrast, being older at the second police contact and having female help both decrease the predicted log odds.

Severity

The model for severity of future crime based on the second police contact is as follows:

$$\begin{aligned}\hat{L} &= -1.032 - 0.211(CURVICSC) + 0.839(FIREARM) + 0.002(KNIFE) + \\ &\quad 0.787(OTHWEAP) + 0.38(FIRSERSC) + 0.505(OFDRRACE) + \\ &\quad 0.083(CURCFAMF) - 0.486(CURCBUSI) - 0.873(CURCPOLI) - \\ &\quad 0.421(CURCSTRA) - 0.041(CURINSID) + 0.424(CUROUTSI)\end{aligned}$$

In this model, higher scores on the victimization scale actually predict lower log odds of falling into the upper quartile of future criminal severity. Having any of the three types of weapons also increases the predicted log odds. The seriousness of one's first criminal offense is a significant predictor of the severity of future offenses, in the logical direction. Among the types of complainants, if the complainant is a family member or friend this tends to increase the log odds slightly, whereas business, police, and strangers as complainants tend to decrease the log odds. Finally, crimes committed indoors decrease the log odds, whereas those committed outside increase them.

1.5 Methods and Results: Model Validation and Diagnostics

To assess the fit and appropriateness of my models, I looked at the following diagnostic measures:

1. $1 - \frac{\text{model deviance}}{\text{null deviance}}$, an R^2 analogue for logistic regression models
2. Expected vs. actual numbers of juveniles in each quintile of risk, for the original data, tested using the χ^2 statistic
3. Expected vs. actual numbers of juveniles in each quintile of risk, for the held-out data, tested using the χ^2 statistic
4. Diagnostic plots using the Pearson and deviance residuals

Please see the Appendix, Section 1.7.7, for computational details.

Quantile	Expected	Observed
1	19.380	12
2	38.638	40
3	54.007	68
4	75.116	73
5	122.875	117
Total	310.017	310
	$\chi^2 = 6.825, p = 0.078$	

Table 1.1: Activity, Original Data

Quantile	Expected	Observed
1	6.134	7
2	12.203	11
3	17.250	19
4	23.829	22
5	40.551	31
Total	99.965	90
	$\chi^2 = 2.808, p = 0.422$	

Table 1.2: Activity, Held-Out Data

1.5.1 Models for the First Contact

The R^2 analogues for the models for activity and severity of future crime at the first police contact were poor, 0.076 and 0.027 respectively, indicating that although a number of variables are significant predictors in each case, the models still do a poor job of predicting whether an individual will go on to be a serious offender. These low values are not due to having removed insignificant variables from the model chosen by the stepwise procedure; for instance, for the model for activity chosen by stepwise the value is still only 0.090.

In order to gauge the fit of the model in a different way, I calculated the predicted probability of falling into the top 25% of offenders based on each model and compared this to the observed probability. Specifically, I looked at these expected and observed probabilities within each quantile of the fitted probabilities generated by each model. Tables 1.1 and 1.2 show expected and observed probabilities for activity, using the fitted values from the data originally used to fit the model and the fitted values found by the model predictions on the held-out data. The accompanying χ^2 statistic and associated p-value test the null hypothesis that the model provides an adequate fit.

Although the fit looks questionable on the original data, the model actually does a fairly good job of allocating the right number of juveniles to each quintile of risk when used for prediction on the new data. By inspection of the table, it

Quantile	Expected	Observed
1	56.463	52
2	76.361	87
3	112.237	113
4	89.120	88
5	133.854	128
Total	468.036	468
$\chi^2 = 2.110, p = 0.550$		

Table 1.3: Severity, Original Data

Quantile	Expected	Observed
1	18.064	21
2	22.903	22
3	29.462	37
4	33.545	34
5	42.668	32
Total		
$\chi^2 = 5.115, p = 0.164$		

Table 1.4: Severity, Held-Out Data

looks like the model tended to over-predict the number of juveniles that should be placed into the highest two quintiles of risk.

Similar tables for the models relating to severity are Tables 1.3 and 1.4. The fit of the model appears to be adequate both for the original data and for the new data.

For the model for activity based on the first police contact, I made two of the diagnostic plots suggested by Hosmer and Lemeshow (1989). Cases in the dataset are said to have the same *X pattern* if they have identical X variables. Diagnostic plots for the residuals focus not on individual cases but on patterns (because these cases have identical fitted probabilities). If we define $\Delta\chi^2_P(j)$ to be the decrease in the sum of squared Pearson residuals that results from deleting all cases with the jth X pattern, and $\Delta\chi^2_D(j)$ to be the change in deviance (sum of squared deviance residuals) that results from deleting all cases with the jth X pattern, then two useful graphs are to plot $\Delta\chi^2_P(j)$ and $\Delta\chi^2_D(j)$ against the predicted probabilities. Higher values indicate poorly fit X patterns. Thus, in Figure 1.3, the labeled points represent patterns whose predicted probability of being in the top quartile of offenders was low, and yet who did in fact go on to be serious offenders. These patterns were characterized by scoring "0" for all or most of those variables which tended to increase the log odds, and to be older, which tended to decrease the log odds. Therefore, their predicted probabilities were very low.

I did not create similar graphs for severity or for the models based on the second contact because the function I used to find the delta values (which is computationally intensive) can currently only handle up to seven covariates, and the other models have more.

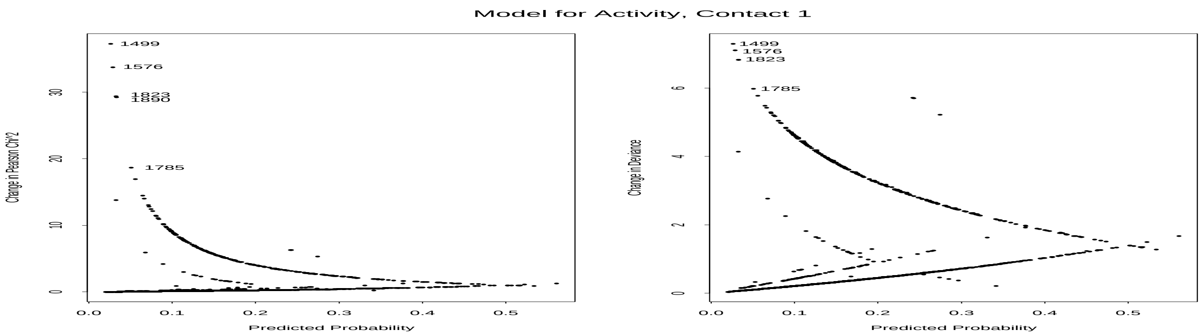

Figure 1.4: Change in χ^2 Values Vs. Fitted Probabilities for Model Predicting Activity at First Contact

1.5.2 Models for the Second Contact

The R^2 analogues for the models at the second contact were 0.103 for the activity model and 0.042 for the severity model. Note that there is slightly more available information at the second contact because all juveniles have some (limited) information about at least one prior contact in addition to the current one.

However, the predictions based on these models are not necessarily better than those at the first contact. The model for activity is significantly under-predicting the number of juveniles to be serious offenders in the held-out dataset, in almost all quintiles of predicted risk.

The same is also true of the model for severity. Both models do provide an adequate fit to the original data, however.

Quantile	Expected	Observed
1	13.839	17
2	26.773	27
3	38.335	33
4	52.778	61
5	88.287	82
Total	220.012	220
$\chi^2 = 3.195, p = 0.363$		

Table 1.5: Activity, Original Data

Quantile	Expected	Observed
1	4.184	9
2	8.452	18
3	11.895	15
4	16.363	25
5	27.077	25
Total		
$\chi^2 = 21.860, p < 0.001$		

Table 1.6: Activity, Held-Out Data

Quantile	Expected	Observed
1	30.268	30
2	40.596	39
3	48.522	43
4	58.353	66
5	83.271	83
Total	261.010	260
$\chi^2 = 1.697, p = 0.638$		

Table 1.7: Severity, Original Data

Quantile	Expected	Observed
1	11.359	18
2	13.307	21
3	16.297	20
4	18.975	18
5	27.167	19
Total		
	$\chi^2 = 11.677, p = 0.009$	

Table 1.8: Severity, Held-Out Data

1.6 Discussion and conclusions

Current policy assigns juveniles to HOU prosecution based upon the accumulation of a certain number of serious crimes. This seems reasonable, as the number of prior events was a significant predictor of future offending in three out of four models. However, given that the accuracy of the model predictions was generally poor, one must ask whether this criteria alone (only one out of several predictors) can even come close to accurately identifying those youths most at risk for becoming serious offenders in the future. Taking other covariates from the model into account, such as age at first and second police contact, could potentially improve the predictions, although the R^2 analogues for these models suggest that many juveniles would still be misclassified.

Race was a significant predictor in all four models; Black juveniles tended to have greater odds of being serious criminals as adults. Although selection criteria for the HOU cannot be based on this criteria, knowledge of this effect can influence policy indirectly, in that programs should address those factors that might put Black juveniles more at risk; only a rough estimate of SES was included in this analysis.

The models presented here might finally be used as the basis of a cost-benefit analysis of the HOU program. By examining the fitted probabilities and the actual criminal outcomes of these juveniles, one might approximate the number of "correct" and "incorrect" targets in the current program, and see if the cost of using priority prosecution to the "incorrect" juveniles is balanced by the benefit of reaching the "correct" targets, for example.

1.6.1 Future Work

One of the major weaknesses of the current analysis was that the full richness of the dataset could not be used, because individuals were not identified by number at each contact. This is one possible explanation for why even the best models were poorly fit in this model. The fact that the diagnostic plots for the model predicting activity at the first contact indicated that the poorly fit points tended to have low fitted probabilities and yet be serious criminals as adults supports the hypothesis that additional variables for each individual might improve the fit of the model.

Given that some of the variables in the dataset were constant for individuals, such as age at first event and SES, it might be feasible to attempt to match individuals at different contacts. Using logical rules, such as that a juvenile cannot be older at his first contact than at his second, one might be able to further reduce the ambiguity regarding which cases should be matched. Based on some very preliminary attempts to do this, I expect the process would be relatively labor intensive, and would likely still result in a number of non-unique cases at each contact.

The variables in the dataset were almost exclusively concerned with concrete aspects of a juvenile's current and past offenses. Some knowledge of environmental and psychological variables might improve the models. However, policy is unlikely to adopt selection criteria based upon such variables, and it would be difficult to obtain such information about this cohort. Studies could perhaps be conducted for a more current cohort, in a city which does not currently have a Habitual Offender Unit, but it seems more immediately advisable to make more of the data that we have by attempting to track individuals throughout police contacts.

1.7 Appendix: Analysis Log, Splus Commands

1.7.1 Loading the Data

There are 12 datasets associated with this study. However, there is a great deal of redundancy in the datasets. Referring to the table below, note that:

1. The last four datasets are pooled versions of the first eight, with some variables removed.
2. The study researchers have divided the original data into model construction and model validation subsets.
3. The even and odd numbered datasets are exact replicates of each other, except that odd datasets include variables indicating whether juveniles fell into the top 25% for each of the outcome variables, whereas the even datasets include variables indicating whether juveniles fell into the top 10%.

Because I wanted to randomly select my own model construction and model validation subsets, and because I planned to disregard the 20 dichotomized outcome variables, I used only the files DS1, DS3, DS5, and DS7. I merged these into one file and read the data into Splus, using the original variable names given in the codebook.

File Name	Race of Juveniles	Sample	Dichotomy
DS1	Black	Validation	25/75
DS2	Black	Validation	10/90
DS3	Black	Construction	25/75
DS4	Black	Construction	10/90
DS5	White	Validation	25/75
DS6	White	Validation	10/90
DS7	White	Construction	25/75
DS8	White	Construction	10/90
DS9	Pooled	Validation	25/75
DS10	Pooled	Validation	10/90
DS11	Pooled	Construction	25/75
DS12	Pooled	Construction	10/90

1.7.2 Univariate Analyses of Predictor Variables

The merged dataset contains 115 predictor variables and one variable indicating the police contact number (first, second, and so on). Many of these variables are indicator variables relating to the same construct. Below are univariate statistics for the predictor variables, grouped by category.

Transition number (sequential police contact)

```
>table(juv$UP3TRANS)
   1    2   3   4   5   6   7   8
 2684 1374 873 481 336 244 178 132
```

The following univariate analyses summarize across all police contacts. An interesting side project would be to see how these univariate distributions change across police contact, but to do that is outside the scope of the current project. Instead, in the sections below in which I describe predictive models for the first and second police contact, I examine the by-contact univariate distributions of only those variables which are found to be predictive at each contact.

Current offense: seriousness, charges, etc.

- *Statutory categories*

  The variables indicating which statutory categories a juvenile's acts fell into are one example of dichotomous variables in the dataset which, although they relate to the same concept, are not mutually exclusive. For example, a juvenile's crime could include both a robbery and a serious assault. I wrote a function, f.categorize, to summarize the number of cases in the dataset which had particular combinations of each category, thus creating mutually exclusive sets. For instance, below we see that the most common situation for juveniles in the dataset was to have committed acts that could be categorized as "Serious property crimes," "Other property crimes," and "Other crimes." The next most common occurrence, accounting for 592 juveniles, was to be charged with drug offenses only. The f.categorize function is at the end of this appendix, in section 1.7.8.

  ```
  > f.categorize(juv[,2:10], names.cat=c("Ser Aslt", "NonSer Aslt", "Rob",
                 "Sex", "Weap", "Drug", "Ser Prop", "Oth Prop", "Other"))
  1792 : Ser Prop/Oth Prop/Other"
  592 : Drug"
  537 : Ser Prop/Other"
  458 : Ser Prop/Oth Prop"
  446 : Ser Prop"
  293 : Weap"
  266 : Rob/Ser Prop/Oth Prop/Other"
  137 : NonSer Aslt/Rob/Ser Prop/Oth Prop/Other"
  120 : Ser Aslt/NonSer Aslt/Other"
  115 : Weap/Other"
  # ... many combinations with less than 100 people
  ```

  The dataset contains a variable, NUMCCAT, which according to the codebook records the number of statutory categories. However, because this did not match a variable created by simply summing the category variables, it was excluded from future analysis. Instead, the simple sum was used. The most common combinations involved three separate categories.

```
# Number of statutory categories for current offense
> table(juv$NUMCCAT)
    1    2    3    4   5
 1436 1037 1482 1504 843

# Variable created by summing
> CURNUMCA <- apply(juv[,2:10],1,sum)
> table(CURNUMCA)
    1    2    3   4   5
 1502 1473 2350 643 334
```

- *Seriousness - two measures*

  I did not want to use all nine dichotomous variables above as predictors; instead, I created a summary variable to indicate how serious the current crime was, based on the statutory categories it fell into. Specifically, I gave the following scores to each category: Other crimes = 1, Other property crimes = 2, Serious property crimes = 3, Drug offenses = 4, Weapons violations = 5, Sex crimes = 6, Robbery = 7, Nonserious assaults = 8, and Serious assaults = 9. I then assigned to each juvenile the maximum seriousness score of his particular categories. The overwhelming majority of cases in the dataset have serious property crimes (level 3) as the most serious category. No juveniles have level 2, other (nonserious) property crimes, as their most serious category, which suggests that other property crimes tend to occur in the presence of more serious crimes.

```
> CURMXSER <- rep(NA, dim(juv)[1])
> CURMXSER <- ifelse(juv$CURCC9==1,1,CURMXSER) # Other crimes
> CURMXSER <- ifelse(juv$CURCC8==1,2,CURMXSER) # Other property crimes
> CURMXSER <- ifelse(juv$CURCC7==1,3,CURMXSER) # Serious property crimes
> CURMXSER <- ifelse(juv$CURCC6==1,4,CURMXSER) # Drug offenses
> CURMXSER <- ifelse(juv$CURCC5==1,5,CURMXSER) # Weapons violations
> CURMXSER <- ifelse(juv$CURCC4==1,6,CURMXSER) # Sex crimes
> CURMXSER <- ifelse(juv$CURCC3==1,7,CURMXSER) # Robbery
> CURMXSER <- ifelse(juv$CURCC2==1,8,CURMXSER) # Nonserious assaults
> CURMXSER <- ifelse(juv$CURCC1==1,9,CURMXSER) # Serious assaults
> table(CURMXSER)
   1    3   4   5  6   7   8   9
  40 3233 657 466 75 523 553 755
```

The dataset also contains a variable, SERCURU, which indicates the seriousness of the current offense, using an unknown scale. A boxplot of this variable is shown in Figure 1.5. The distribution of scores is centered around 9, but there are many high outliers.

```
> summary(juv$SERCURU)
  Min. 1st Qu. Median  Mean 3rd Qu.  Max.
  1.00  3.90    8.50   9.28 12.00   87.30
```

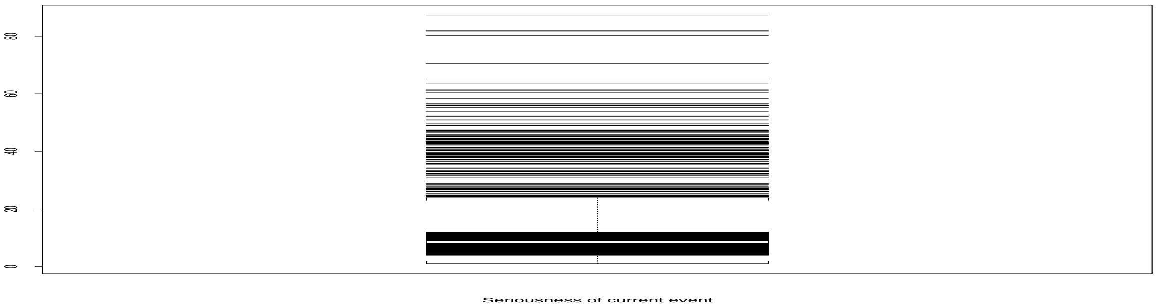

Figure 1.5: Boxplot of variable SERCURU (Seriousness of current offense)

Because parallel boxplots of the original SERCURU variable split by the derived CURMXSER variable (Figure 1.6) did not show a clear relationship between the two, I included both in further analyses.

- *Number of charges*

  Over all police contacts, about 63% resulted in two to four charges being brought against the juvenile. It was less common to receive only one charge or more than five charges.

```
> f.categorize(juv[,83:85], names.cat=c("one only", "two to four", "five+"))
[1] "3865 : two to four"
[1] "1362 : one only"
[1] "1075 : five+"
```

- *Arrest*

  Almost 90% of police contacts resulted in arrest.

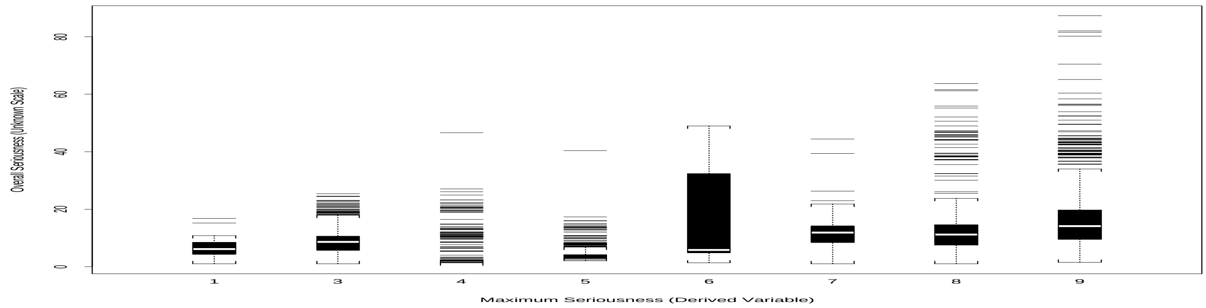

Figure 1.6: Two Measures of Seriousness

```
> table(juv[,90]) # 0=Not arrested , 1=Arrested
   0    1
 747 5555
```

Victims of current offense

- *Whether or not there was a victim*

  The variable NOVICTIM indicates whether or not there was a victim of the juvenile's current crime. Trying different combinations of statutory categories in order to determine what criteria might have been used to create this variable, I could find none that allowed me to recreate it. Because other variables in the dataset were coded as missing based on the coding of this variable, I decided to take it at face value, although I renamed it to have a more natural interpretation.

```
> table(juv$NOVICTIM) # 0=no victim, 1=victim(s)
    0    1
 2746 3556
> CURVICTM <- juv$NOVICTIM==1 # 0=no victim, 1=victim(s)
```

- *Number of victims*

I used the f.categorize function to examine the number of victims, looking first at those cases which CURVICTM indicated did have a victim, and then looking at the entire dataset. The numbers agree. (Note that the f.categorize function indicates those cases in which all dichotomized variables are "0" as "None.")

In the entire dataset, most contacts involved crimes against one victim or no victim. Having multiple victims was much less common.

```
> f.categorize(juv[CURVICTM,55:57],    # Cases in which there was a victim
               names.cat=c("one only", "two or more", "unknown"))
[1] "2973 : one only"
[1] "318 : two or more"
[1] "265 : unknown"
> f.categorize(juv[,55:57],            # All cases
               names.cat=c("one only", "two or more", "unknown"))
[1] "2973 : one only"
[1] "2746 : None"
[1] "318 : two or more"
[1] "265 : unknown"
```

- *Type of victimization*

  Another series of variables indicate the level of victimization that occurred. In order to simplify further analysis, I created one variable to summarize the level of victimization that occurred. Higher scores indicate higher levels of victimization.

```
> f.categorize(juv[,78:82], names.cat=c("no or mutual", "face to face",
                                         "primary", "secondary", "tertiary"))
[1] "2334 : primary"
[1] "1978 : face to face"
[1] "681 : secondary"
[1] "655 : tertiary"
[1] "654 : no or mutual"
> CURVICSC <- rep(NA, dim(juv)[1])
> CURVICSC <- ifelse(juv$TYPEVIC1==1,0,CURVICSC) # No or mutual
> CURVICSC <- ifelse(juv$TYPEVIC5==1,1,CURVICSC) # Tertiary
> CURVICSC <- ifelse(juv$TYPEVIC4==1,2,CURVICSC) # Secondary
> CURVICSC <- ifelse(juv$TYPEVIC3==1,3,CURVICSC) # Primary
> CURVICSC <- ifelse(juv$TYPEVIC2==1,4,CURVICSC) # Face-to-face
> table(CURVICSC)
   0   1   2    3    4
 654 655 681 2334 1978
```

- *Bodily injury*

The variable CURSW1 indicates whether or not the current crime resulted in bodily injury, although the codebook does not specify who specifically was injured. About 20% of contacts were associated with some injury.

```
> table(juv$CURSW1)
   0    1
5026 1276
```

- *Level of injury inflicted on victim(s)*

  Another series of variables indicates how much the victim specifically was injured. I again looked only at those cases in which CURVICTM indicated there was a victim as well as the entire dataset. The numbers agree because cases with no victim are also coded as having no injury to the victim. Finally, I created another scale variable to indicated the degree of injury to the victim. Higher scores indicate higher levels of injury.

```
> f.categorize(juv[CURVICTM,106:110], names.cat=c("death or hospitalization",
               "treated or discharged", "minor harm", "no injury", "unknown"))
[1] "2233 : no injury"
[1] "614 : minor harm"
[1] "368 : treated or discharged"
[1] "198 : death or hospitalization"
[1] "143 : unknown"
> f.categorize(juv[,106:110], names.cat=c("death or hospitalization",
               "treated or discharged", "minor harm", "no injury", "unknown"))
[1] "4979 : no injury"
[1] "614 : minor harm"
[1] "368 : treated or discharged"
[1] "198 : death or hospitalization"
[1] "143 : unknown"
> CURINJSC <- rep(NA, dim(juv)[1])
> CURINJSC <- ifelse(juv$CURSW1D==1,0,CURINJSC) # no injury"
> CURINJSC <- ifelse(juv$CURSW1C==1,1,CURINJSC) # minor harm
> CURINJSC <- ifelse(juv$CURSW1B==1,2,CURINJSC) # treated or discharged
> CURINJSC <- ifelse(juv$CURSW1A==1,3,CURINJSC) # death or hospitalization
> table(CURINJSC)
   0   1   2   3
4979 614 368 198
```

- *Intimidation*

```
Three indicator variables deal with the type of intimidation, if any,
used by the juvenile.  Most did not use physical intimidation or
intimidation with a weapon, although they may have used verbal intimidation.
```

```
> f.categorize(juv[,45:47],
               names.cat=c("phys", "weap", "none, unknown, or verbal"))
[1] "5400 : none, unknown, or verbal"
[1] "467 : phys"
[1] "435 : weap"
```

- *Age of victim(s)*

  The variable VICTAGEU had been set to the mean age when the age of the victim was unknown:

```
> table(juv$VICTAGE[CURVICTM&juv$VICTAGEU==1]) # Victim but unknown age
 32.5313
     956
```

  I created a new variable for age, changing these unknown cases to NA. If there was no victim the age was also set to NA. This variable is summarized below, and a histogram of the ages appears in Figure 1.7.

```
> CURVIAGE <- rep(NA, dim(juv)[1])
> CURVIAGE[CURVICTM&juv$VICTAGEU==0] <- juv$VICTAGE[CURVICTM&juv$VICTAGEU==0]
> summary(CURVIAGE)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   1.00   16.00   28.00   32.11   45.00   97.00 3702.00
```

- *Gender of victim(s)*

  Where there is at least one known victim, the distribution of victim gender is as follows. Note that over twice as many males as females were victims.

```
> f.categorize(juv[CURVICTM,51:54],        # Cases where there was a victim
               names.cat=c("males only", "females only", "mixed", "unknown"))
[1] "2131 : males only"
[1] "947 : females only"
[1] "406 : unknown"
[1] "72 : mixed"
```

- *Race of victim(s)*

  The races of victims were fairly evenly divided between white and nonwhite victims. It was rare for a crime to involve both white and nonwhite victims, although we saw above that it was relatively rare for crimes to involve more than one victim, period.

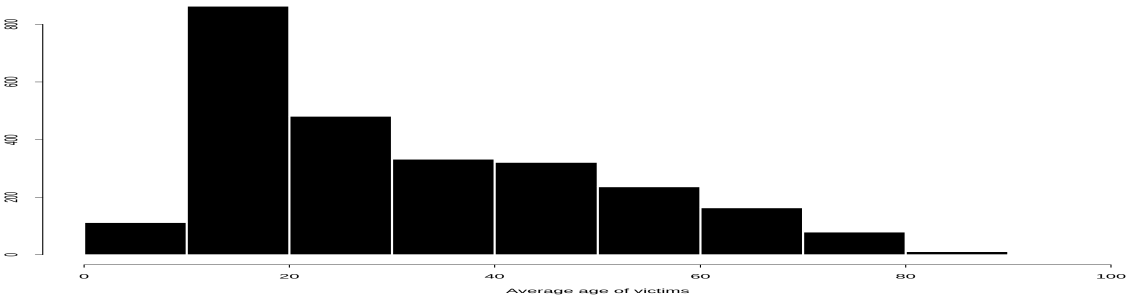

Figure 1.7: Histogram of variable CURVIAGE (Average age of victims)

```
> f.categorize(juv[CURVICTM,70:73], names.cat=c("whites only", "nonwhites only",
                                                 "mixed", "unknown"))
[1] "1593 : nonwhites only"
[1] "1265 : whites only"
[1] "682 : unknown"
[1] "16 : mixed"
```

Weapons used in current offense

There are two groups of variables which code the type of weapon used in the current crime. One group has to do with firearms, and the other with knives, although there is some overlap. I used the f.categorize function to determine where the overlap occurred:

```
> f.categorize(juv[,c(26:28, 102:105)],
             names.cat=c("firearms", "something other than firearms",
                         "no or unknown weapon", "knife",
                         "something other than a knife", "none",
                         "unknown"))
[1] "4968 : no or unknown weapon"                      # NONE
[1] "580 : something other than firearms/knife"        # KNIFE
[1] "402 : firearms"                                   # FIREARMS
[1] "323 : something other than firearms"              # OTHER
[1] "21 : firearms/something other than firearms/knife"  # FIREARMS + KNIFE
```

```
[1] "8 : firearms/something other than firearms"        # FIREARMS + OTHER
```

Given the mutually exclusive categories listed by f.categorize, I created four new indicator variables for the four types of weapons. Using f.categorize on these variables led to the same results, but now we have four variables to deal with instead of seven:

```
> CURWEAP1 <- ifelse(juv$WEAPCURA==1,1,0)                # FIREARMS
> CURWEAP2 <- ifelse(juv$KNIFCURA==1,1,0)                # KNIFE
> CURWEAP3 <- ifelse(juv$WEAPCURB==1&juv$KNIFCURA==0,1,0)  # OTHER
> CURWEAP4 <- ifelse(juv$WEAPCURC==1,1,0)                # NONE OR UNKNOWN
> f.categorize(data.frame(cbind(CURWEAP1, CURWEAP2, CURWEAP3,CURWEAP4)),
               names.cat=c("Firearms", "Knife", "Other", "None or unknown"))
[1] "4968 : None or unknown"
[1] "580 : Knife"
[1] "402 : Firearms"
[1] "323 : Other"
[1] "21 : Firearms/Knife"
[1] "8 : Firearms/Other"
```

Offenders in current offense

- *Number of offenders*

  The majority of crimes in the dataset were committed by more than one offender.

  ```
  > f.categorize(juv[,32:34], names.cat=c("one", "two+", "unknown"))
  [1] "3342 : two+"
  [1] "2283 : one"
  [1] "677 : unknown"
  ```

- *Gender of offenders*

  Although all juveniles in the dataset are male, sometimes they had help from female offenders. However, this occurred in less than 2% of the known cases.

  ```
  > f.categorize(juv[,35:37], names.cat=c("males", "mixed", "unknown"))
  [1] "5498 : males"
  [1] "700 : unknown"
  [1] "104 : mixed"
  ```

- *Age of offenders*

  Again, although all juveniles in the dataset were less than 18 years old at each recorded police contact, they sometimes had adult help. This occurred in about 10% of the known cases.

```
> f.categorize(juv[,42:44], names.cat=c("juv", "juv+adult", "unknown"))
[1] "4965 : juv"        # JUVENILES ONLY
[1] "770 : unknown"
[1] "567 : juv+adult"    # HAD ADULT HELP
```

Setting of current offense

- *Type of location*

  Most crimes occurred in public places, although relatively large numbers of crimes also occurred in business and residential locations.

```
> f.categorize(juv[,58:61],
               names.cat=c("business or commercial", "residence",
                           "public place", "unknown"))
[1] "3841 : public place"
[1] "1037 : business or commercial"
[1] "757 : residence"
[1] "665 : unknown"
[1] "2 : None"
> # Two don't have anything coded for site -- recode to unknown
> nosite <- apply(juv[,58:61],1,sum)==0
> juv[nosite,]
     UP3TRANS CURCC1 CURCC2 CURCC3 CURCC4 CURCC5 CURCC6
3084        5      0      0      0      0      0      0
5373        1      0      0      0      0      0      0
 ...
> juv[,61] <- juv[,61]+nosite # Change 0 to 1 for "unknown"
> f.categorize(juv[,58:61],
               names.cat=c("business or commercial", "residence",
                           "public place", "unknown"))
[1] "3841 : public place"
[1] "1037 : business or commercial"
[1] "757 : residence"
[1] "667 : unknown"
```

- *Environment (indoor vs. outdoor)*

  Likewise, about 58% of known cases occurred outside instead of inside.

```
> f.categorize(juv[,62:64], names.cat=c("inside", "outside", "unknown"))
[1] "3277 : outside"
[1] "2405 : inside"
[1] "620 : unknown"
```

Police response to current offense

- *Complainant's relationship to offender*

  Police response was brought about by complaints; these were most often from strangers or businesses, although sometimes the police themselves were the complainants.

```
> f.categorize(juv[,65:69], names.cat=c("family or friend", "business",
                                        "police", "stranger", "unknown"))
[1] "2062 : stranger"
[1] "1452 : business"
[1] "1213 : police"
[1] "850 : unknown"
[1] "725 : family or friend"
```

- *Reason for police response*

  Similar to the last set of variables, these indicate why the police responded to a given crime. Usually they received a complaint or observed the crime themselves.

```
> f.categorize(juv[,86:89], names.cat=c("resp to complaint", "police suspicion",
                                        "police observed ofs", "unknown"))
[1] "3790 : resp to complaint"
[1] "953 : police observed ofs"
[1] "858 : unknown"
[1] "701 : police suspicion"
```

Prior events

- *Number of prior events*

  Most contacts in the dataset had been preceded by two or more offenses.

```
> f.categorize(juv[,91:93], names.cat=c("none", "one", "two or more"))
[1] "3678 : two or more"
[1] "1586 : none"
[1] "1038 : one"
```

- *Statutory categories*

  The indicator variables for the statutory categories of past offenses had been set to the mean in cases where there were structural zeros. For this analysis I reset these cases to zero. Note that the number of juveniles categorized as "None" now agrees with the number of juveniles who had no prior events above.

```
> juv$ANYCC2 <- ifelse(juv$ANYCC2==1, 1, 0)
> juv$ANYCC8 <- ifelse(juv$ANYCC8==1, 1, 0)
> juv$ANYCC9 <- ifelse(juv$ANYCC9==1, 1, 0)
> f.categorize(juv[,11:19], names.cat=c("Ser Aslt", "NonSer Aslt", "Rob",
              "Sex", "Weap", "Drug", "Ser Prop", "Oth Prop", "Other"))
[1] "1586 : None"
[1] "964 : Other"
[1] "930 : Ser Prop/Oth Prop/Other"
[1] "217 : Rob/Ser Prop/Oth Prop/Other"
[1] "205 : Ser Prop/Other"
[1] "184 : NonSer Aslt/Rob/Ser Prop/Oth Prop/Other"
[1] "134 : NonSer Aslt/Ser Prop/Oth Prop/Other"
[1] "120 : Ser Aslt/NonSer Aslt/Rob/Ser Prop/Oth Prop/Other"
[1] "112 : NonSer Aslt/Other"
[1] "110 : Ser Aslt/NonSer Aslt/Rob/Weap/Ser Prop/Oth Prop/Other"
[1] "110 : Weap/Ser Prop/Oth Prop/Other"
> # ... many combinations with less than 100 people
```

- *Seriousness*

  Just as I did for the current offense, I created a variable to designate the maximum seriousness of the categories for past offenses. If there were no prior offenses, this was coded as zero. Increasing scores indicate greater maximum seriousness.

```
> PRIMXSER <- rep(NA, dim(juv)[1])
> PRIMXSER <- ifelse(juv$ANYCC9==1,1,PRIMXSER) # Other crimes
> PRIMXSER <- ifelse(juv$ANYCC8==1,2,PRIMXSER) # Other property crimes
> PRIMXSER <- ifelse(juv$ANYCC7==1,3,PRIMXSER) # Serious property crimes
> PRIMXSER <- ifelse(juv$ANYCC6==1,4,PRIMXSER) # Drug offenses
> PRIMXSER <- ifelse(juv$ANYCC5==1,5,PRIMXSER) # Weapons violations
> PRIMXSER <- ifelse(juv$ANYCC4==1,6,PRIMXSER) # Sex crimes
> PRIMXSER <- ifelse(juv$ANYCC3==1,7,PRIMXSER) # Robbery
> PRIMXSER <- ifelse(juv$ANYCC2==1,8,PRIMXSER) # Nonserious assaults
> PRIMXSER <- ifelse(juv$ANYCC1==1,9,PRIMXSER) # Serious assaults
> PRIMXSER <- ifelse(is.na(PRIMXSER),0,PRIMXSER) # No prior events
> table(PRIMXSER)
    0   1  2    3   4   5  6   7   8   9
 1586 964 12 1211 208 263 58 349 898 753
```

- *Legal disposition of prior events*

  Before examining the legal responses to prior events, I first reset structural zeros back to zero (instead of the mean of known cases). 1586 juveniles had no legal response, which is because they had committed no prior crimes. There were a large number of unknown cases.

```
> juv$EVTDISP1 <- ifelse(juv$EVTDISP1==1, 1, 0)
> juv$EVTDISP2 <- ifelse(juv$EVTDISP2==1, 1, 0)
> juv$EVTDISP3 <- ifelse(juv$EVTDISP3==1, 1, 0)
> juv$EVTDISP4 <- ifelse(juv$EVTDISP4==1, 1, 0)
> f.categorize(juv[,21:24], names.cat=c("Adjud, confined",
             "Adjud, not confined", "Unknown", "Not adjud"))
[1] "1586 : None"
[1] "1195 : Unknown"
[1] "1119 : Unknown/Not adjud"
[1] "736 : Adjud, not confined/Unknown/Not adjud"
[1] "645 : Not adjud"
[1] "349 : Adjud, not confined/Not adjud"
[1] "184 : Adjud, confined/Adjud, not confined/Unknown/Not adjud"
[1] "149 : Adjud, not confined/Unknown"
[1] "144 : Adjud, not confined"
[1] "106 : Adjud, confined/Unknown/Not adjud"
[1] "29 : Adjud, confined/Not adjud"
[1] "20 : Adjud, confined/Adjud, not confined/Not adjud"
[1] "20 : Adjud, confined/Unknown"
[1] "15 : Adjud, confined/Adjud, not confined/Unknown"
[1] "4 : Adjud, confined"
[1] "1 : Adjud, confined/Adjud, not confined"
```

- *Specific prior crimes*

  Again, I first reset structural zeros to zero. The following tables summarize the number of contacts in the dataset in which the juvenile was known to have committed specific prior crimes. The most common were prior nuisance crimes and prior status violations.

```
> juv$PRIGANG <- ifelse(juv$PRIGANG==1,1,0)
> juv$PRISTAT <- ifelse(juv$PRISTAT==1,1,0)
> juv$PRILIQ <- ifelse(juv$PRILIQ==1,1,0)
> juv$PRIDRUNK <- ifelse(juv$PRIDRUNK==1,1,0)
> juv$PRIDISOR <- ifelse(juv$PRIDISOR==1,1,0)
> juv$PRISOLV <- ifelse(juv$PRISOLV==1,1,0)
> juv$PRIVAND <- ifelse(juv$PRIVAND==1,1,0)
> juv$PRINUIS <- ifelse(juv$PRINUIS==1,1,0)
> table(juv$PRIGANG)    #Gang related crimes
   0   1
 5902 400
> table(juv$PRISTAT)    #Status violations
   0    1
 3891 2411
> table(juv$PRILIQ)     #Liquor law violations
   0   1
```

```
 6186 116
> table(juv$PRIDRUNK)   #Public drunkenness
    0   1
 6157 145
> table(juv$PRIDISOR)   #Disorderly conduct
    0    1
 4583 1719
> table(juv$PRISOLV)    #Solvent use
    0  1
 6204 98
> table(juv$PRIVAND)    #Prior vandalism
    0    1
 5085 1217
> table(juv$PRINUIS)    #Prior nuisance crimes
    0    1
 3838 2464
```

The above tables do not account for the fact that a given juvenile may have committed more than one prior crime. Here are the top combinations of prior crimes:

```
> f.categorize(juv[,94:101], names.cat=c("Gang", "Status", "Liquor", "Drunk",
               "Disorderly", "Solvent", "Vandalism", "Nuisance"))
[1] "2623 : None"
[1] "1107 : Status"
[1] "482 : Disorderly/Nuisance"
[1] "382 : Status/Disorderly/Nuisance"
[1] "315 : Status/Vandalism/Nuisance"
[1] "271 : Vandalism/Nuisance"
[1] "246 : Status/Disorderly/Vandalism/Nuisance"
[1] "179 : Disorderly/Vandalism/Nuisance"
> # ... many combinations with less than 100 people
```

- *Seriousness of earliest prior event and most recent prior event*

  These two variables involve the same scale used for the current event. Boxplots for these two variables are in Figure 1.8.

```
> summary(juv$SERFEVTU) # Earliest prior event
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  0.000  0.900   1.411  3.521  4.300  47.100
> summary(juv$SERMEVTU)  # Most recent prior event
    Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  0.000  1.167   2.300  4.406  5.697  61.200
```

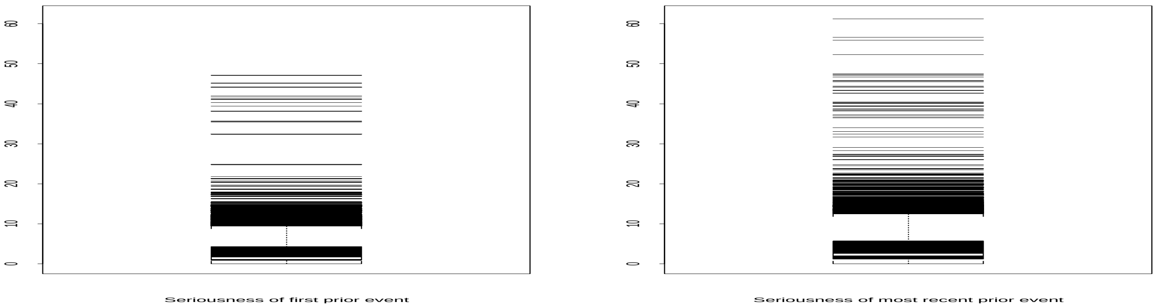

Figure 1.8: Boxplots of variables SERFEVTU and SERMEVTU (Seriousness of first prior and most recent prior events)

Personal characteristics of the offender

- *SES of offender*

  Most contacts were with juveniles in the middle range of socio-economic status. Few were in the top 15%.

```
> f.categorize(juv[,74:76], names.cat=c("top 15%", "midrange", "bottom 15%"))
[1] "4235 : midrange"
[1] "1767 : bottom 15%"
[1] "300 : top 15%"
```

- *Race of offender*

  The dataset contained contacts with 1178 white juveniles and 5124 black juveniles.

```
 # 1=black, 0=white
> table(juv[,77])
     0    1
  1178 5124
```

- *Offender age at current offense*

Parallel boxplots of offender age by police contact number are in Figure 1.9. Note that the upper limit of the distribution is always 18 because all contacts are during the juvenile years, that median age increases with police contact (from about 15.5 at first contact to 16.5 at eighth contact), and that there are many low outliers.

```
> summary(juv$AGECURU)
    Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  5.451 14.610  15.960 15.510 16.870  18.000
```

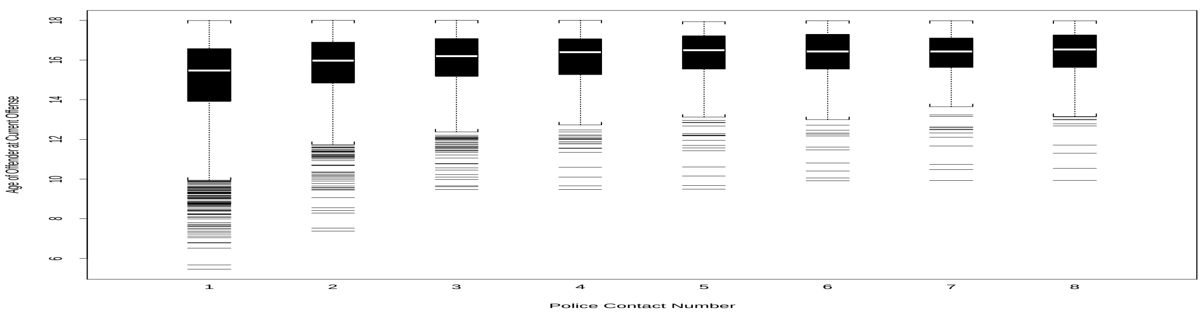

Figure 1.9: Boxplots of offender age by police contact number (1st through 8th contact)

- *Offender age at earliest prior event and at most recent prior event*

  You can see in Figure 1.10 the distributions of these variables by police contact, for the first three contacts. Clearly something is going on with the first contact; it seems that in cases in which there was no prior event, these variables have been set to their means. However, because these cases make up such a large portion of the data for the first and second police contact, resetting these variables to be NA would eliminate many cases from a straightforward analysis. Instead, I did not include these two variables as covariates in any further analysis.

```
> # Age at first prior event
```

```
> summary(juv$AGEFEVTU)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  2.259 12.470  13.230 13.030 14.120  17.910
> # Age at most recent event
> summary(juv$AGEMEVTU)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  7.307 14.600  14.840 15.080 15.940  17.970
> #Time since most recent event
> summary(juv$AGECURU-juv$AGEMEVTU)
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -9.1670  0.0466  0.3531  0.4297  1.1630  8.8730
> # Indicates problems -- doesn't make sense that negative time has elapsed
> # another reason to remove these variables
```

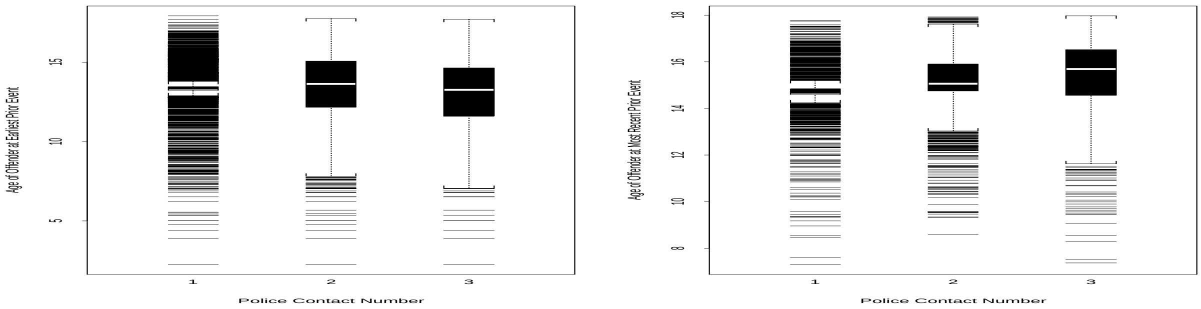

Figure 1.10: Boxplots of offender age at earliest and most recent prior events, police contacts 1 to 3

1.7.3 Outcome Variables

There were 40 continuous outcome variables in the dataset; 20 of them dealt with crimes committed from the current contact to the individual's 18th birthday, and 20 of them dealt with crimes committed from the current contact to the individual's 27th birthday. I only dealt with this latter set of variables.

```
> Y <- juv[,seq(113,189,4)]
```

```
> names(Y)
 [1] "EVTPYAUC" "UCRPYAUC" "VIOPYAUC" "UP3PYAUC" "EVTSPAUC" "UCRSPAUC"
 [7] "VIOSPAUC" "UP3SPAUC" "EVTSYAUC" "UCRSYAUC" "VIOSYAUC" "UP3SYAUC"
[13] "EVTCPAUC" "UCRCPAUC" "VIOCPAUC" "UP3CPAUC" "EVTCYAUC" "UCRCYAUC"
[19] "VIOCYAUC" "UP3CYAUC"
```

There was a considerable amount of colliniarity among the outcome variables:

```
> f.collin(Y)
EVTPYAUC 0.99
UCRPYAUC 0.995
VIOPYAUC 0.989
UP3PYAUC 0.996
EVTSPAUC 0.921
UCRSPAUC 0.959
VIOSPAUC 0.915
UP3SPAUC 0.961
EVTSYAUC 0.994
UCRSYAUC 0.996
VIOSYAUC 0.986
UP3SYAUC 0.997
EVTCPAUC 0.928
UCRCPAUC 0.961
VIOCPAUC 0.874
UP3CPAUC 0.967
EVTCYAUC 0.999
UCRCYAUC 0.999
VIOCYAUC 0.993
UP3CYAUC 0.999
```

I ran a principal components analysis on the outcome variables in order to determine their structure. I did this instead of arbitrarily choosing outcome variables from among the twenty. You can see the factor loadings above .2 below. The first two factors have clear interpretations – 1 is mainly related to the per year variables (I called this component "activity"), and 2 is mainly related to the per contact variables (I called this component "severity"). Note that the various types of crime, indicated by the 3 letter prefix for each variable, had approximately equal loadings. Although the third factor had an eigen-value over 1, I used only the first two factors in my analyses for ease of interpretation. These factors accounted for 83% of the overall variance; see the scree plot in Figure 1.11. A biplot of the original data and the 20 variables is in Figure 1.12.

```
> pc.cor <- princomp(Y, cor=T)
> print(loadings(pc.cor), cutoff=.2) # Only the first 8 out of 20 are shown
         Comp. 1 Comp. 2 Comp. 3 Comp. 4 Comp. 5 Comp. 6 Comp. 7 Comp. 8
EVTPYAUC  0.237                   0.230   0.244           -0.397
```

```
UCRPYAUC  0.249                                                   0.428
VIOPYAUC  0.233                  -0.427  -0.204          -0.255
UP3PYAUC  0.249
EVTSPAUC         -0.348           0.221  -0.341   0.496  -0.311
UCRSPAUC         -0.369                          -0.412           0.238
VIOSPAUC         -0.295  -0.391           0.481           0.257   0.252
UP3SPAUC         -0.372  -0.203   0.227                  -0.356
EVTSYAUC  0.261                                                  -0.454
UCRSYAUC  0.262                                           0.403
VIOSYAUC  0.234          -0.372  -0.320                   0.231  -0.281
UP3SYAUC  0.262                                           0.209  -0.351
EVTCPAUC         -0.266   0.383          -0.284   0.393   0.254
UCRCPAUC         -0.278   0.380                  -0.380          -0.213
VIOCPAUC                         -0.372   0.592   0.327  -0.218
UP3CPAUC         -0.283   0.397                  -0.208
EVTCYAUC  0.257
UCRCYAUC  0.256                                                   0.330
VIOCYAUC  0.234                  -0.439  -0.205          -0.206
UP3CYAUC  0.258
> eigen(cor(Y))$values[1:5]
[1] 13.1070569  3.6624595  1.0709519  0.7855224  0.5672696
```

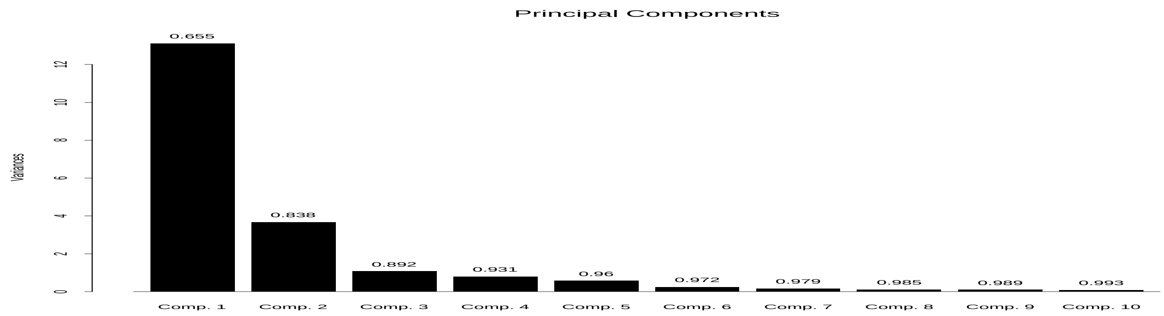

Figure 1.11: Scree plot of principal components

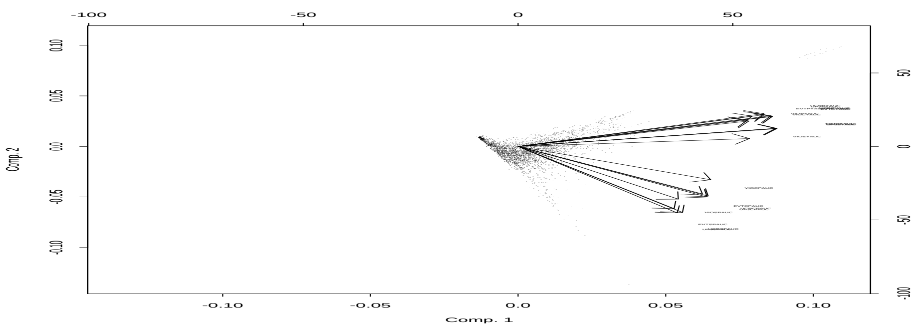

Figure 1.12: Biplot of principal components

I created two new variables from the component scores, reversing the sign of the second component scores so that both variables would have a similar interpretation, with higher values indicating greater crime. Histograms of the distributions of these variables are in Figure 1.13.

```
> activity <- pc.cor$scores[,1]
> summary(activity)
     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  -4.0040  -3.1630  -0.3712   0.0000   2.0440  31.4600
> severity <- -pc.cor$scores[,2]
> summary(severity)
      Min.    1st Qu.    Median      Mean   3rd Qu.      Max.
 -15.10000  -1.45000  -0.07083   0.00000   1.02400  20.63000
```

Noting that the factor loadings above suggested that the principal component scores were approximately the mean of the related variables, I created two new variables corresponding to these means.

```
> mean.activity <- (juv$EVTPYAUC + juv$UCRPYAUC + juv$VIOPYAUC + juv$UP3PYAUC +
+                   juv$EVTSYAUC + juv$UCRSYAUC + juv$VIOSYAUC + juv$UP3SYAUC +
+                   juv$EVTCYAUC + juv$UCRCYAUC + juv$VIOCYAUC + juv$UP3CYAUC)/12
> mean.severity <- (juv$EVTSPAUC+juv$UCRSPAUC+juv$VIOSPAUC+juv$UP3SPAUC+
+                   juv$EVTCPAUC+juv$UCRCPAUC+juv$VIOCPAUC+juv$UP3CPAUC)/8
```

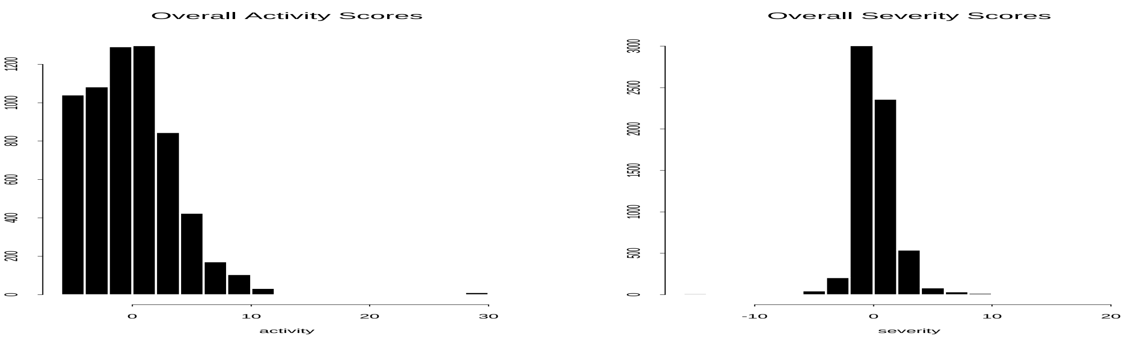

Figure 1.13: New outcome variables

Plotting these two mean variables against their respective principal component scores (Figure 1.14), the means for activity were a good approximation to the first principal component scores, but the means for severity were not as a good an approximation to the second principal component scores. Therefore, I decided to use the principal component scores themselves for the analysis. However, one should bear in mind that the interpretation of the outcome variables is similar to simply taking the mean of the related variables.

Finally, because the research objective was to be able to predict the worst future offenders, I dichotomized the two outcome variables, activity and severity, by splitting the data at the 75th percentile. Thus, the operational definition of a serious offender was someone who fell into the top 25% of all offenders. Actually, this is not quite precise, since each case represents a police contact, not an individual (many individuals have repeat contacts). However, doing a separate principal components analysis and dichotomization at each contact would not be appropriate either, because then the meaning of the outcome variables would be that an individual was in the top 25% of offenders, *among those juveniles who had a greater or equal number of police contacts.*

Here are the changing numbers of juveniles categorized as being in the top 25% for activity and severity, across police contacts:

```
> table(act.dum, juv$UP3TRANS)
     1    2   3   4   5   6  7  8
0 2284 1062 612 297 184 129 93 65
```

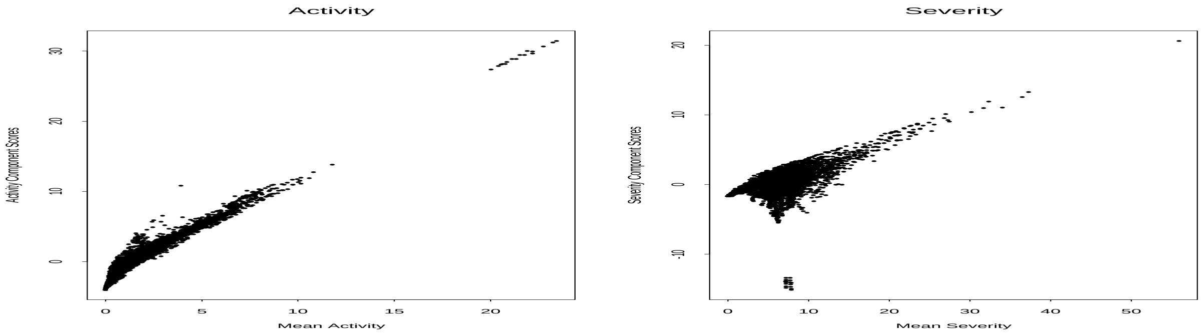

Figure 1.14: Mean Variables Vs. Principal Component Scores

```
1  400   312 261 184 152 115 85 67
> table(sev.dum, juv$UP3TRANS)
     1    2   3   4   5   6   7   8
0 2070 1017 641 345 245 177 128 103
1  614  357 232 136  91  67  50  29
```

1.7.4 Final Data for Analysis

For reference, following are two tables of the variables that were used to build predictive models for frequency and severity at the first and second police contact. They roughly correspond to the variables discussed above, although some variables relating to victims were omitted due to great amounts of missing data. Baseline categories indicate what it means if an individual variable, or all related variables, are zero.

1.7.5 Predictive Models: First Police Contact

I selected the data for the first police contacts, holding out a random subset of 1/4 of the data:

```
crime1 <- crime[crime$UP3TRANS==1,]
crime1.hold <- crime1[samp1,]
crime1 <- crime1[-samp1,]
```

Topic	Variable(s)	Original Variable	Interpretation	Baseline Category (if applicable)
	UP3TRANS	UP3TRANS	Police contact number	
Current offense	CURNUMCA	see (1.7.2)	Number of statutory categories	
	CURMXSER	see (1.7.2)	Max seriousness of categories	
	CURSERSC	SERCURU	Seriousness (unknown scale)	
	CURCH2.4 CURCH5..	NUMCHRG2 NUMCHRG3	2-4 current charges 5+ current charges	1 charge only
	CURARRST	ARRDUM	Was arrested	Not arrested
Victims of current offense	CURVICTM	NOVICTIM	Had a victim	No victim
	CURVICSC	(see 1.7.2)	Extent of victimization (0-4)	
	CURINJUR	CURSW1	Was bodily injury	No injury
	CURINJSC	(see 1.7.2)	Extent of injury to victim (0-3)	No injury
	CURINTWP	INTIMB2	Intimidation w/ weapon	No
	CURINTPH	INTIMB1	Physical intimidation	No
Weapons in current offense	CURWEAP1 CURWEAP2 CURWEAP3	(see 1.7.2) (see 1.7.2) (see 1.7.2)	Had firearm Had knife Had other weapon	No No No
Offenders in current offense	CURMULOF	OFDRNUMB	More than one offender	No or unknown
	CURFEMOF	OFDRGENB	Female help	No or unknown
	CURADTOF	JUVADTB	Adult help	No or unknown
Setting for current offense	CURSITE1 CURSITE2 CURSITE3	SITEA SITEB SITEC	Commercial site Residential site Public site	Unknown
	CURINSID CUROUTSI	INOUTA INOUTB	Inside Outside	Unknown
Response to current offense	CURCFAMF CURCBUSI CURCPOLI CURCSTRA	COMREL2A COMREL2B COMREL2C COMREL2D	Complainant family/friend Complainant business Complainant police Complainant stranger	Unknown
	CURRCOMP CURRSUSP CURROBSR	POLRESP1 POLRESP2 POLRESP3	Police responded to complaint Police responded to suspicion Police responded to observation	Unknown

Topic	Variable(s)	Original Variable	Interpretation	Baseline Category (if applicable)
Prior events	PRIEVNT1 PRIEVNT2	PRIEVT1 PIEVT2	One prior event 2+ prior events	No prior events
	PRIMXSER	(see 1.7.2)	Max seriousness of prior categories	
	PRIDISP1 PRIDISP2 PRIDISP3	EVTDISP1 EVTDISP2 EVTDISP3	Adjudicated, confined Adjudicated, not confined Not adjudicated	Unknown
	PRIGANGR	PRIGANG	Gang related crimes	No
	PRISTATV	PRISTAT	Status violations	No
	PRILIQLV	PRILIQ	Liquor law violations	No
	PRIDRUNK	PRIDRUNK	Public drunkenness	No
	PRIDISOR	PRIDISOR	Disorderly conduct	No
	PRISOLVU	PRISOLV	Solvent use	No
	PRIVANDL	PRIVAND	Prior vandalism	No
	PRINUISC	PRINUIS	Pubic nuisance crimes	No
	FIRSERSC	SERFEVTU	Seriousness of first offense (unknown scale)	
Personal characteristics	OFHIHSES OFLOWSES	SESCATA SESCATC	In top 15% SES In bottom 15% SES	Middle 15% SES
	OFDRRACE	RACEVAR	African American	White
	OFNDRAGE	AGECURU	Age at current offense	
	OFFIRAGE	AGEFEVTU	Age at first offense	
Outcomes	act.dum	(see 1.7.3)	In top 25& of crime activity	Bottom 75%
	sev.dum	(see 1.7.3)	In top 25& of crime severity	Bottom 75%
	activity	(see 1.7.3)	Activity component scores	
	severity	(see 1.7.3)	Severity component scores	

Because I had so many predictor variables, I used the following strategy to choose models for predicting the top 25% of offenders in terms of the two principal components, activity and severity of future crime:

1. First I used the glm() function in Splus to fit a logistic model using all predictor variables.
2. I applied the step.glm() function to the glm object above in order to determine the best model from among all variables (without interactions), based on AIC.
3. I then added back in any variables that were related to those chosen by the stepwise function.
4. Finally, I iteratively removed variables which were deemed insignificant by the likelihood ratio test by arranging to put them last in the model and then using anova(model, test="Chisq").

This produced voluminous output which is available upon request. Here, I summarize the models chosen by the stepwise function and the final models for each of the outcome variables.

Activity

```
> # Model chosen by stepwise:
> summary(crime1.act.step)$coef
                  Value Std. Error   t value
(Intercept) -0.64451807 0.62611386 -1.029394
   CURMXSER  0.12517263 0.04910482  2.549090
   CURSERSC  0.02872749 0.01209288  2.375572
   CURARRST  0.40146038 0.20682349  1.941077
   CURINJSC -0.28576568 0.16140448 -1.770494
   CURINTWP -0.50344095 0.33982961 -1.481451
   CURINTPH -0.48533138 0.31002134 -1.565477
   CURFEMOF -0.96683465 0.73081247 -1.322959
   PRIEVNT1  0.50534184 0.16745049  3.017858
   PRIEVNT2  0.97017299 0.17645014  5.498284
   PRIVANDL  0.47649001 0.21342701  2.232567
   OFDRRACE  1.01489734 0.19266812  5.267593
   OFNDRAGE -0.16399972 0.03426913 -4.785639
   OFFIRAGE -0.06727097 0.04020985 -1.672997
> # Final model:
> summary(crime1.act.final)

Call: glm(formula = act.dum ~ CURARRST + PRIEVNT1 + PRIEVNT2 + PRIVANDL + OFDRRACE +
        OFNDRAGE, family = binomial, data = crime1.act, na.action = na.omit)
Deviance Residuals:
       Min         1Q     Median         3Q      Max
```

```
 -1.282878 -0.6247211 -0.4767565 -0.3066639 2.699278

Coefficients:
                 Value Std. Error   t value
(Intercept) -0.9761219 0.48281033 -2.021750
   CURARRST  0.5232185 0.19601741  2.669245
   PRIEVNT1  0.4904383 0.16389360  2.992419
   PRIEVNT2  1.0134269 0.16968685  5.972336
   PRIVANDL  0.5158473 0.20568948  2.507894
   OFDRRACE  1.0112575 0.18649036  5.422572
   OFNDRAGE -0.1628795 0.02977785 -5.469821

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 1729.503 on 2012 degrees of freedom

Residual Deviance: 1597.321 on 2006 degrees of freedom

Number of Fisher Scoring Iterations: 4

Correlation of Coefficients:
         (Intercept)   CURARRST   PRIEVNT1   PRIEVNT2   PRIVANDL   OFDRRACE
CURARRST -0.2410245
PRIEVNT1  0.0520870  -0.0533105
PRIEVNT2  0.1013780  -0.0542906  0.4361753
PRIVANDL -0.0682219  -0.0133725 -0.2084835 -0.3945252
OFDRRACE -0.3936622  -0.0378679 -0.0404691 -0.0175405  0.1115546
OFNDRAGE -0.8607956  -0.1012888 -0.1529068 -0.2137548  0.0406813  0.0813863
```

Severity

```
> # Model chosen by stepwise:
> summary(crime1.sev.step)$coefficient
                 Value Std. Error    t value
(Intercept) -2.1605785 0.20740842 -10.417024
   CURNUMCA  0.1054886 0.06484404   1.626804
   CURCH2.4 -0.3286913 0.11958520  -2.748595
   CURVICSC  0.1203200 0.05614797   2.142909
   CURINTWP -0.5681891 0.27863803  -2.039166
   CURWEAP2  0.3389914 0.18899130   1.793688
   CURADTOF  0.2681132 0.19043533   1.407896
   CUROUTSI  0.2795395 0.11455950   2.440125
   CURRCOMP -0.2228261 0.13369562  -1.666667
   CURROBSR -0.3148177 0.18322899  -1.718165
   PRIEVNT1  0.2257265 0.13533434   1.667918
   PRIEVNT2  0.4697303 0.13983770   3.359110
```

```
   PRILIQLV -1.3736924 1.05052218  -1.307628
   PRIDRUNK -1.6039917 1.01689292  -1.577346
   OFDRRACE  0.5897909 0.14153301   4.167161
> # Final model:
> summary(crime1.sev.final)

Call: glm(formula = sev.dum ~ CURVICSC + OFDRRACE + PRIEVNT1 + PRIEVNT2 + CURCH2.4 +
        CURCH5.. + CURINSID + CUROUTSI, family = binomial, data = crime1.sev,
        na.action = na.omit)
Deviance Residuals:
       Min         1Q     Median         3Q      Max
 -1.094112 -0.7685154 -0.6576017 -0.4700248 2.174341

Coefficients:
                  Value Std. Error    t value
(Intercept) -1.95834009 0.22758746 -8.6047804
   CURVICSC  0.11775572 0.04903873  2.4012801
   OFDRRACE  0.60819037 0.13573834  4.4806086
   PRIEVNT1  0.18251285 0.13225620  1.3799947
   PRIEVNT2  0.38242584 0.13525727  2.8273958
   CURCH2.4 -0.16550101 0.14720282 -1.1243060
   CURCH5..  0.25140557 0.20191204  1.2451242
   CURINSID -0.25901818 0.18147435 -1.4272992
   CUROUTSI  0.04618668 0.17217193  0.2682591

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 2183.164 on 2012 degrees of freedom

Residual Deviance: 2124.749 on 2004 degrees of freedom

Number of Fisher Scoring Iterations: 3

Correlation of Coefficients:
         (Intercept)   CURVICSC   OFDRRACE   PRIEVNT1   PRIEVNT2   CURCH2.4
CURVICSC -0.3537588
OFDRRACE -0.4461792  -0.0864305
PRIEVNT1 -0.1230595   0.0085265 -0.0395089
PRIEVNT2 -0.1118575   0.0034370  0.0074931  0.2968261
CURCH2.4 -0.1527356  -0.4781986  0.0005815 -0.0244656 -0.0439133
CURCH5.. -0.0324364  -0.4983759 -0.0004261 -0.0247636 -0.0801910  0.6453655
CURINSID -0.6085685   0.1380954 -0.0071758 -0.0272266 -0.0478048 -0.0837186
CUROUTSI -0.6575193   0.1028906  0.0547489 -0.0422954 -0.0599369 -0.0586692

           CURCH5..   CURINSID
CURVICSC
```

```
OFDRRACE
PRIEVNT1
PRIEVNT2
CURCH2.4
CURCH5..
CURINSID -0.1001526
CUROUTSI -0.0803953  0.7802373
```

1.7.6 Predictive Models: Second Police Contact

I followed the same procedure as for the first contact to find models based on the second contact. However, I removed the variable PRIEVNT1 (the indicator for having one prior event) because at the second contact it was impossible to have no prior events. Thus, including PRIEVNT1 and PRIEVNT2 would have been redundant. If PRIEVENT2 is coded as 1, then the juvenile had two or more prior events. If it is coded as zero, then he had only one.

Activity

```
> # Model chosen by stepwise:
> summary(crime2.act.step)$coef
                  Value Std. Error   t value
(Intercept)  0.81942631 0.84575295  0.968872
   CURSERSC -0.01910014 0.01276411 -1.496394
   CURARRST  0.86349244 0.32746306  2.636916
   CURVICTM  0.54550653 0.23527218  2.318619
   CURWEAP2  0.69735715 0.26957274  2.586898
   CURFEMOF -5.78924471 5.40607999 -1.070877
   CURADTOF -0.46424845 0.38482497 -1.206389
   CURSITE1 -0.64684465 0.55730552 -1.160664
   CURSITE2 -0.64880924 0.58043944 -1.117790
   CURSITE3 -0.70921594 0.53359075 -1.329138
   CURINSID  0.62074579 0.54520605  1.138553
   CUROUTSI  0.81974401 0.55919271  1.465942
   CURCPOLI  0.32446753 0.28310274  1.146112
   CURCSTRA -0.34417860 0.23574533 -1.459959
   CURROBSR -0.31937488 0.25494693 -1.252711
   PRIEVNT2  0.57304758 0.24777794  2.312747
   PRIMXSER  0.04698616 0.03665167  1.281965
   PRIDISP1  1.64182398 0.46000179  3.569169
   PRIDISP3  0.44539180 0.24266047  1.835453
   PRIGANGR  0.67337981 0.36337164  1.853144
   PRISTATV -0.24604680 0.20784228 -1.183815
   PRILIQLV  1.18740593 0.57360162  2.070088
   PRIVANDL  0.37843245 0.20954031  1.806013
   OFDRRACE  0.70547201 0.25003647  2.821476
```

```
   OFNDRAGE -0.35201425 0.06353849 -5.540174
   OFFIRAGE  0.05727504 0.04992402  1.147244
> # Final model
> summary(crime2.act.final)

Call: glm(formula = act.dum ~ CURARRST + CURWEAP1 + CURWEAP2 + CURWEAP3 + PRIEVNT2 +
        PRIDISP1 + PRIDISP2 + PRIDISP3 + OFDRRACE + OFNDRAGE + CURFEMOF, family
         = binomial, data = crime2.act, na.action = na.omit)
Deviance Residuals:
       Min         1Q     Median        3Q      Max
 -1.717013 -0.7098605 -0.5363395 -0.281176 2.521072

Coefficients:
                  Value Std. Error    t value
(Intercept)  1.46038546 0.78399947  1.8627378
   CURARRST  0.81223096 0.31660814  2.5654140
   CURWEAP1  0.22925273 0.33021846  0.6942456
   CURWEAP2  0.71141537 0.25709603  2.7671192
   CURWEAP3  0.09537143 0.34664668  0.2751258
   PRIEVNT2  0.54753410 0.21620232  2.5325080
   PRIDISP1  1.61153022 0.44868807  3.5916493
   PRIDISP2  0.19123332 0.19047624  1.0039746
   PRIDISP3  0.23674497 0.22177822  1.0674852
   OFDRRACE  0.69278375 0.23607135  2.9346372
   OFNDRAGE -0.31240743 0.04752269 -6.5738589
   CURFEMOF -5.75846619 5.57227439 -1.0334140

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 1068.475 on 1029 degrees of freedom

Residual Deviance: 958.4074 on 1018 degrees of freedom

Number of Fisher Scoring Iterations: 6

Correlation of Coefficients:
         (Intercept)  CURARRST   CURWEAP1   CURWEAP2   CURWEAP3   PRIEVNT2
CURARRST -0.2268563
CURWEAP1  0.0435005  -0.0347515
CURWEAP2  0.0394881   0.0890156  0.0981333
CURWEAP3  0.0115273   0.0585602  0.0782850  0.1086977
PRIEVNT2  0.1219361   0.0500416 -0.0357194  0.0104580  0.0089847
PRIDISP1  0.0394252  -0.0234975 -0.0149499 -0.0210786  0.0054286 -0.0340966
PRIDISP2  0.0539208   0.0216888 -0.0435044 -0.0322967 -0.0160624 -0.0926292
PRIDISP3 -0.3585428  -0.0649825  0.0092743 -0.0235856 -0.0078288 -0.5007355
OFDRRACE -0.2476244  -0.0591564 -0.0788420 -0.0898148 -0.0428206 -0.0013122
```

```
OFNDRAGE -0.8635252  -0.1450693 -0.0378261 -0.0950430 -0.0583653 -0.2527016
CURFEMOF -0.0001635   0.0027607  0.0016036 -0.0025209  0.0036414 -0.0005019

           PRIDISP1   PRIDISP2   PRIDISP3   OFDRRACE   OFNDRAGE
CURARRST
CURWEAP1
CURWEAP2
CURWEAP3
PRIEVNT2
PRIDISP1
PRIDISP2  0.0610389
PRIDISP3  0.0261248  0.1605580
OFDRRACE  0.0605511 -0.0084571  0.0072271
OFNDRAGE -0.0727201 -0.1483083  0.2844890  0.0162197
CURFEMOF  0.0026754 -0.0053669 -0.0051570  0.0028067 -0.0017813
```

Severity

```
> # Model chosen by stepwise:
> summary(crime2.sev.step)$coefficient
                  Value Std. Error   t value
(Intercept) -1.39620327 0.41283599 -3.381981
   CURVICSC -0.23128883 0.08296436 -2.787810
   CURWEAP1  0.91470883 0.30179487  3.030896
   CURWEAP3  0.82154464 0.29197736  2.813727
   CURSITE1  1.32953865 0.50976040  2.608164
   CURSITE2  1.16229323 0.49287947  2.358169
   CURSITE3  1.34826601 0.42623963  3.163164
   CURINSID -0.37557878 0.22668269 -1.656848
   CURCBUSI -0.61312202 0.33549522 -1.827513
   CURCPOLI -0.86185295 0.35461256 -2.430407
   CURCSTRA -0.46788293 0.23140517 -2.021921
   CURRCOMP -0.53836674 0.32575893 -1.652654
   CURRSUSP -0.82490569 0.40808627 -2.021400
   CURROBSR -0.57335560 0.37090573 -1.545826
   PRIDISP1  0.74952092 0.43619783  1.718305
   PRISOLVU  0.93823397 0.53929681  1.739736
   FIRSERSC  0.04102166 0.01112670  3.686775
   OFDRRACE  0.55201803 0.20351133  2.712468
> # Final model:
> summary(crime2.sev.final)

Call: glm(formula = sev.dum ~ CURVICSC + CURWEAP1 + CURWEAP2 + CURWEAP3 + FIRSERSC +
        OFDRRACE + CURCFAMF + CURCBUSI + CURCPOLI + CURCSTRA + CURINSID +
        CUROUTSI, family = binomial, data = crime2.sev, na.action = na.omit)
Deviance Residuals:
```

```
      Min         1Q     Median        3Q      Max
 -1.500257 -0.7753278 -0.6649318 0.9885958 2.165651

Coefficients:
                   Value Std. Error      t value
(Intercept) -1.032244921 0.38087409 -2.710199896
   CURVICSC -0.211151623 0.07982307 -2.645245418
   CURWEAP1  0.838668802 0.29948273  2.800391210
   CURWEAP2  0.001699641 0.27185608  0.006251989
   CURWEAP3  0.787453731 0.28933850  2.721565691
   FIRSERSC  0.038481209 0.01097098  3.507546448
   OFDRRACE  0.505023038 0.19784779  2.552583609
   CURCFAMF  0.083127514 0.36735740  0.226285124
   CURCBUSI -0.486318465 0.35576983 -1.366946906
   CURCPOLI -0.872907663 0.38815912 -2.248839739
   CURCSTRA -0.421188798 0.32406389 -1.299709116
   CURINSID -0.041860091 0.37923191 -0.110381245
   CUROUTSI  0.423738119 0.36365164  1.165230879

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 1166.037 on 1029 degrees of freedom

Residual Deviance: 1117.099 on 1017 degrees of freedom

Number of Fisher Scoring Iterations: 3

Correlation of Coefficients:
         (Intercept)  CURVICSC   CURWEAP1   CURWEAP2   CURWEAP3   FIRSERSC
CURVICSC -0.6027005
CURWEAP1  0.1144987  -0.1617389
CURWEAP2  0.1078584  -0.1468726  0.1494742
CURWEAP3  0.0677835  -0.1698535  0.1351245  0.1380168
FIRSERSC -0.1492529  -0.0034091  0.0436378 -0.0008332  0.0541310
OFDRRACE -0.3886973  -0.1007205 -0.0861306 -0.0961923 -0.0060177 -0.0113711
CURCFAMF -0.0589645  -0.0577797 -0.0765154 -0.0747119 -0.0625416  0.0017815
CURCBUSI -0.2011559   0.1804149 -0.0496921 -0.0184452 -0.0453163 -0.0083564
CURCPOLI -0.3702357   0.4553685 -0.1715011 -0.1310006 -0.1363153  0.0155057
CURCSTRA -0.0603486  -0.0554251  0.0071351  0.0086668 -0.0430888 -0.0096513
CURINSID -0.3559625   0.0372090  0.0011363 -0.0142042  0.0148150 -0.0188182
CUROUTSI -0.3477774  -0.0164204 -0.0307797 -0.0396104 -0.0002157 -0.0138630

           OFDRRACE   CURCFAMF   CURCBUSI   CURCPOLI   CURCSTRA   CURINSID
CURVICSC
CURWEAP1
CURWEAP2
```

```
CURWEAP3
FIRSERSC
OFDRRACE
CURCFAMF  0.0310950
CURCBUSI  0.0146705  0.6797322
CURCPOLI  0.0104257  0.5935256  0.7210806
CURCSTRA  0.0220987  0.7278922  0.7526790  0.6864718
CURINSID -0.0002819 -0.5715941 -0.6613588 -0.5047591 -0.6253061
CUROUTSI  0.0263142 -0.5412425 -0.5778463 -0.5709260 -0.6708038  0.8794979
```

1.7.7 Model Diagnostics and Interpretation

R^2 Analogues

I calculated an R^2 analogue for each of my models, for predicting activity and severity at the first and second contacts.

```
> # Activity, contact 1
> 1 - crime1.act.final$deviance/crime1.act.final$null.deviance
> [1] 0.07642766
>
> # Severity, contact 1
> 1 - crime1.sev.final$deviance/crime1.sev.final$null.deviance
> [1] 0.02675676
>
> # Activity, contact 2
> 1 - crime2.freq.final$deviance/crime2.freq.final$null.deviance
> [1] 0.1030141
>
> # Severity, contact 2
> 1 - crime2.sev.final$deviance/crime2.sev.final$null.deviance
> [1] 0.04197012
```

Because these numbers were so low, I wanted to see what the actual R^2 might have been, using the continuous versions of the outcome variables and regressing on all possible predictors. Here are the results for the first contact:

```
> allvars <- lm(activity ~ ., data=crime1.act, na.action=na.omit)
> summary(allvars)$r.squared
> [1] 0.1648284
>
> allvars <- lm(severity ~ ., data=crime1.sev, na.action=na.omit)
> summary(allvars)$r.squared
[1] 0.05567163
```

So the largest possible values of R^2 (without using interactions, etc.) indicate that it would be difficult to get good predictions using the data as it stands, analyzing at each contact. The predictive value of the dataset would potentially be much richer if we could track individuals through the dataset, across contacts, but to try to identify match up the cases that represent individuals would be difficult and time-consuming, considering that there are only a few non-changing variables in the dataset (such as age at first contact, race, and SES), and many of these are the same across many individuals or are missing.

Diagnostic Plots

Here is the Splus code for the diagnostic plots mentioned in Section 1.5.1. This makes use of Howard Seltman's "logitinf" function to extract the values of $\Delta\chi^2_{P(j)}$, $\Delta\chi^2_{D(j)}$, the unique patterns among the X variables, etc.

```
> crime1.freq.final.data <- as.data.frame(cbind(crime1$CURARRST, crime1$PRIEVNT1,
+                                                crime1$PRIEVNT2, crime1$PRIVANDL,
+                                                crime1$OFDRRACE, crime1$OFNDRAGE))
> inf <- logitinf(crime1.freq.final, crime1.freq.final.data, crime1$freq.dum)
> # Distinct Patterns
> dim(inf$coll)[1] # [1] 1886
> # Original data
> dim(crime1)[1] # [1] 2013
>
> # Pearson
> par(mfrow=c(1,2))
> plot(inf$coll$fit, inf$DelChiP, xlab="Predicted Probability",
+      ylab="Change in Pearson Chi^2")
> Bad <- identify(inf$coll$fit, inf$DelChiP, inf$coll$pat, cex=1)
>
> # Deviance
> plot(inf$coll$fit, inf$DelChiD, xlab="Predicted Probability",
+      ylab="Change in Deviance")
> Bad <- c(Bad, identify(inf$coll$fit, inf$DelChiD, inf$coll$pat, cex=1))
> par(mfrow=c(1,1)); title(main="Model for Activity, Contact 1")
>
> Bad <- unique(Bad)
> round(cbind(inf$coll, DCP=inf$DelChiP, DCD=inf$DelChiD, DB=inf$DelB)[Bad,],2)
    X.1 X.2 X.3 X.4 X.5   X.6 Y m  fit obs hat  pat   DCP  DCD   DB
1499   0   0   0   0   0 16.21 1 1 0.03   1   0 1499 37.27 7.30 0.06
1576   0   0   0   0   0 15.60 1 1 0.03   1   0 1576 33.75 7.11 0.06
1823   1   0   0   0   0 17.96 1 1 0.03   1   0 1823 29.38 6.83 0.04
1890   1   0   0   0   0 17.93 1 1 0.03   1   0 1890 29.23 6.83 0.04
1785   0   1   0   0   0 14.97 1 1 0.05   1   0 1785 18.69 5.97 0.07
```

Predictor Variables: Univariate Distributions and Interpretation

Here are the univariate distributions for the relevant predictor variables for each contact:

Contact 1:

```
> table(crime1$CURARRST) # More common to be arrested than not
   0    1
 356 1657
> f.categorize(crime1[,31:32], names.cat=c("1", "2+")) # Majority have no prior events
[1] "1173 : None"
[1] "449 : 1"
[1] "391 : 2+"
> table(crime1$PRIVANDL) # Majority have not vandalized
    0   1
 1842 171
> table(crime1$OFDRRACE) # Majority are African American
   0    1
 517 1496
> summary(crime1.freq$OFNDRAGE) # Mean is about 15
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  5.451 13.950  15.470 15.040 16.560  18.000
> table(crime1$CURVICSC) # Level of victimization varies; often high
   0   1   2   3   4
 257 244 255 725 532
> f.categorize(crime1[,4:5], names.cat=c("2-4", "5+")) # Most 2-4 charges
[1] "1253 : 2-4"
[1] "506 : None"
[1] "254 : 5+"
> f.categorize(crime1[,22:23], names.cat=c("Inside", "Outside")) # Most outside
[1] "1038 : Outside"
[1] "749 : Inside"
[1] "226 : None"
```

Contact 2:

```
> table(crime2$CURARRST) # Most arrested
   0   1
 114 916
> f.categorize(crime2[,13:15], names.cat=c("1","2","3")) # Most have no weapon
[1] "822 : None"
[1] "89 : 2"
[1] "61 : 3"
[1] "57 : 1"
[1] "1 : 1/2"
> table(crime2$PRIEVNT2) # Most have two prior events
   0   1
 332 698
> f.categorize(crime2[,33:35], names.cat=c("1","2","3")) # Most not adjudicated
[1] "526 : 3"
[1] "211 : None"
```

```
[1] "139 : 2/3"
[1] "130 : 2"
[1] "13 : 1/3"
[1] "7 : 1"
[1] "4 : 1/2/3"
> table(crime2$OFDRRACE) # Most African American
   0   1
 217 813
> summary(crime2$OFNDRAGE)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  7.373 14.810  15.950 15.630 16.930  18.000
> table(crime2$CURFEMOF) # Rare to have female help
    0  1
 1015 15
> table(crime2$CURVICSC) # Level of victimization varies, often high
   0   1   2   3   4
 107 103 100 382 338
> summary(crime2$FIRSERSC) # (Seriousness of first offense)
   Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  0.000  0.900   2.850  5.222  8.758  47.100
> f.categorize(crime2[,24:27], names.cat=c("FAMF", "BUSI", "POLI", "STRA"))
[1] "357 : STRA"
[1] "240 : BUSI"
[1] "205 : POLI"
[1] "125 : None"
[1] "103 : FAMF"
> f.categorize(crime2.sev[,22:23], names.cat=c("inside", "outside"))
[1] "528 : outside"
[1] "410 : inside"
[1] "92 : None"
```

For interpretation of the model coefficients, it is useful to exponentiate them. For reference, here are those values:

```
> exp(crime1.act.final$coef)
 (Intercept) CURARRST PRIEVNT1 PRIEVNT2 PRIVANDL OFDRRACE  OFNDRAGE
   0.3767694  1.68745 1.633032 2.755026 1.675057 2.749056 0.8496936
> exp(crime1.sev.final$coef)
 (Intercept) CURVICSC OFDRRACE PRIEVNT1 PRIEVNT2 CURCH2.4 CURCH5.. CURINSID
   0.1410924 1.124969 1.837104  1.20023 1.465836 0.847469 1.285831 0.771809

 CUROUTSI
  1.04727
> exp(crime2.act.final$coef)
 (Intercept) CURARRST CURWEAP1 CURWEAP2 CURWEAP3 PRIEVNT2 PRIDISP1 PRIDISP2
     4.30762 2.252929  1.25766 2.036872 1.100067 1.728984 5.010472 1.210742
```

```
 PRIDISP3 OFDRRACE  OFNDRAGE    CURFEMOF
 1.267118 1.999273 0.7316834 0.003155949
> exp(crime2.sev.final$coef)
 (Intercept)  CURVICSC CURWEAP1 CURWEAP2 CURWEAP3 FIRSERSC OFDRRACE CURCFAMF
   0.3562064 0.8096513 2.313285 1.001701 2.197793 1.039231 1.657024  1.08668

 CURCBUSI  CURCPOLI  CURCSTRA  CURINSID CUROUTSI
 0.614886 0.4177351 0.6562662 0.9590039 1.527661
```

Finally, here is the Splus code used to generate the conditional effects plot.

```
> # Activity, Contact 1:
> # Conditional effect plot for age, grouping by # prior events
> # Group 1: Currently arrested, no prior events, no prior vandalism, black
> # Group 2: Currently arrested, 1 prior event, no prior vandalism, black
> # Group 2: Currently arrested, 2 prior events, no prior vandalism, black
>
> OFNDRAGE <- rep(seq(5,18,.1),3)
> CURARRST <- rep(0,393)
> PRIEVNT1 <- c(rep(0,131), rep(1,131), rep(0,131))
> PRIEVNT2 <- c(rep(0,131), rep(0,131), rep(1,131))
> PRIVANDL <- rep(0,393)
> OFDRRACE <- rep(1,393)
>
> prdata <- as.data.frame(cbind(CURARRST,PRIEVNT1,PRIEVNT2,PRIVANDL,OFDRRACE, OFNDRAGE))
>
> L <- predict(crime1.act.final, prdata)
> P <- 1/(1+exp(-L))
>
> plot(OFNDRAGE, P, xlab="Age at first police contact",
+      ylab="Predicted probabilty of future serious offending", type="n")
>lines(OFNDRAGE[PRIEVNT1==0&PRIEVNT2==0], P[PRIEVNT1==0&PRIEVNT2==0], lty=1)
>lines(OFNDRAGE[PRIEVNT1==1&PRIEVNT2==0], P[PRIEVNT1==1&PRIEVNT2==0], lty=2)
>lines(OFNDRAGE[PRIEVNT1==0&PRIEVNT2==1], P[PRIEVNT1==0&PRIEVNT2==1], lty=3)
> legend(locator(1), c("No prior events", "One prior events",
+                     "Two or more prior events"), lty=1:3)
> title(main="Effect of age on predictions for frequency",
+       sub ="Held constant: Arrest=T, Prior Vandalism=F, Race=African American")
```

Distribution of Predicted Offenders vs. Actual Offenders

In order to gauge the fit of the model, I calculated the predicted probability of falling into the top 25% of offenders based on each model and compared this to the observed probability. Here is the code for one model; the others are similar.

```
>fits <- fitted(crime2.final)
```

```
>
> # Overall
> sum(fits) # Predicted number
[1] 220.0119
> sum(crime2$act.dum) # Actual number
[1] 220
> # Makes sense because the model was fit on this data
>
> summary(fits)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
 0.0002498 0.1149000 0.1874000 0.2136000 0.2795000 0.8011000
>
> quantile(fits, c(.2,.4,.6,.8,1))
      20%       40%       60%       80%      100%
 0.1026381 0.1592969 0.2160839 0.3069207 0.8011271
>
> quintiles <- quantile(fits, c(.2,.4,.6,.8,1))
>
> exp1 <- sum(fits[fits<=quintiles[1]]) # [1] 13.83943
> exp2 <- sum(fits[fits>quintiles[1]&fits<=quintiles[2]]) # [1] 26.77288
> exp3 <- sum(fits[fits>quintiles[2]&fits<=quintiles[3]]) # [1] 38.33492
> exp4 <- sum(fits[fits>quintiles[3]&fits<=quintiles[4]]) # [1] 52.77782
> exp5 <- sum(fits[fits>quintiles[4]&fits<=quintiles[5]]) # [1] 88.28683
>
> obs1 <- sum(crime2$act.dum[fits<=quintiles[1]]) # [1] 17
> obs2 <- sum(crime2$act.dum[fits>quintiles[1]&fits<=quintiles[2]]) # [1] 27
> obs3 <- sum(crime2$act.dum[fits>quintiles[2]&fits<=quintiles[3]]) # [1] 33
> obs4 <- sum(crime2$act.dum[fits>quintiles[3]&fits<=quintiles[4]]) # [1] 61
> obs5 <- sum(crime2$act.dum[fits>quintiles[4]&fits<=quintiles[5]]) # [1] 82
>
> expected <- c(exp1, exp2, exp3, exp4, exp5)
> observed <- c(obs1, obs2, obs3, obs4, obs5)
> terms <- (expected-observed)^2/expected
> chisq <- sum(terms, na.rm=T) # [1] 3.194761
> 1-pchisq(chisq, 3)
[1] 0.3625605
```

Predictions on Held-Out Data

I used a similar method to compare the predictions on the held-out data to the actual values; here some example code.

```
> L <- predict(crime2.freq.final, crime2.hold)
> fits <- 1/(1+exp(-L))
>
> # Overall
```

```
> sum(fits) # Predicted number in held-out data
[1] 67.9705
> sum(crime2.hold$act.dum) # Actual number
[1] 92
> # The model is predicting that more juveniles will
> # be serious offenders than is actually the case.
>
> summary(fits)
      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
 0.0001567 0.1059000 0.1762000 0.1976000 0.2574000 0.7077000
>
> quantile(fits, c(.2,.4,.6,.8,1))
        20%     40%       60%       80%      100%
 0.09671784 0.14741 0.1985228 0.2744638 0.7077419
>
> quintiles <- quantile(fits, c(.2,.4,.6,.8,1))
>
> exp1 <- sum(fits[fits<=quintiles[1]]) # [1] 4.183633
> exp2 <- sum(fits[fits>quintiles[1]&fits<=quintiles[2]]) # [1] 8.451838
> exp3 <- sum(fits[fits>quintiles[2]&fits<=quintiles[3]]) # [1] 11.89546
> exp4 <- sum(fits[fits>quintiles[3]&fits<=quintiles[4]]) # [1] 16.3626
> exp5 <- sum(fits[fits>quintiles[4]&fits<=quintiles[5]]) # [1] 27.07697
>
> obs1 <- sum(crime2.hold$act.dum[fits<=quintiles[1]]) # [1] 9
> obs2 <- sum(crime2.hold$act.dum[fits>quintiles[1]&fits<=quintiles[2]]) # [1] 18
> obs3 <- sum(crime2.hold$act.dum[fits>quintiles[2]&fits<=quintiles[3]]) # [1] 15
> obs4 <- sum(crime2.hold$act.dum[fits>quintiles[3]&fits<=quintiles[4]]) # [1] 25
> obs5 <- sum(crime2.hold$act.dum[fits>quintiles[4]&fits<=quintiles[5]]) # [1] 25
>
> expected <- c(exp1, exp2, exp3, exp4, exp5)
> observed <- c(obs1, obs2, obs3, obs4, obs5)
> terms <- (expected-observed)^2/expected
> chisq <- sum(terms, na.rm=T) # [1] 21.8605
> 1-pchisq(chisq, 3)
[1] 6.97e-05        # Significantly bad!
```

1.7.8 Splus Functions

```
f.bin2dec <- function(x, n=length(x)){  # x is a vector of 1s and 0s
  sum(x*2^((n-1):0))
}

Example:
> x <- c(1,1,0,1,0)
> f.bin2dec(x)
[1] 26
```

```
f.dec2bin <- function(x, n){  # x is a decimal number
  binary <- rep(0,n)
  for (i in 1:n){
    if (x-2^(n-i)>=0){
      binary[i] <- 1
      x <- x-2^(n-i)
    }
  }
  binary
}

Example:
> f.dec2bin(26,5)
[1] 1 1 0 1 0

f.categorize <- function(x.data.frame, names.cat=names(x.data.frame)) {

  num.cat <- length(names(x.data.frame))

  level.table <- sort(table(apply(as.matrix(x.data.frame), 1, f.bin2dec, n=num.cat)))
  counts <- as.numeric(level.table)

  decimals <- as.numeric(names(level.table))

  binaries <- lapply(decimals, f.dec2bin, n=num.cat)

  #Make the table
  for (i in length(counts):1){
    names.i <- names.cat[binaries[[i]]==1]
    if(sum(binaries[[i]])==0) print(paste(counts[i],": None"))
    else print(paste(counts[i],":",paste(names.i, collapse="/")))
  }

  return(NULL)
}
```

Bibliography

[1] Gottfredson, S. D., Warner, B. D., and Taylor, R. B., "Conflict and Consensus in Justice System Decisions," in N. Walker and M. Hough, (eds.), Sentencing and the Public. Cambridge Series in Criminology. London: Gower, 1988.

[2] Hosmer, D. W. and Lemeshow, S. Applied Logistic Regression. New York: Wiley, 1989.

[3] Venables, W. N., and Ripley, B. D., Modern Applied Statistics with S-Plus. New York: Springer, 1999.

[4] Weiner, N. A. Early Identification of the Serious Habitual Juvenile Offender Using a Birth Cohort in Philadelphia, 1958-1984 [Computer file]. ICPSR version. Philadelpha, PA: University of Pennsylvania, 1996. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 1998.