Analysis of wrongly convicted of crimes since 1989

JI LI

1. Introduction

A. Background

Researchers in the United States have identified well over 2,000 individuals who were wrongly convicted of crimes that they did not commit since 1989. We are curious about the compensation of the wrongly convicted. By way of background, there are two basic ways the wrongfully convicted can seek compensation. First, 33 states have state statutes which permit exonerees to request compensation from a state court or state administrative body. Second, the wrongly convicted may instead (or also) seek compensation in court for a violation of their civil rights or on various state law tort theories like false imprisonment or malicious prosecution. Plus, not every exoneree is actually incarcerated and not everyone who is incarcerated seeks compensation.

B. Goals

This report is to better understand the relationship between important factors, such as race, gender, years lost, etc. and the compensation outcome of the wrongly convicted of crimes. I need to respond to some inquiries from Professor Jeffrey Gutman of the GW Law School, fit multivariate regression and make some predictions of unsolved cases.

2. Dataset and data pre-processing

Those individuals and considerable data about each of them are detailed in the National Registry of Exonerations – http://www.law.umich.edu/special/exoneration/Pages/about.aspx. The dataset contains 1900 exonerees organized alphabetically by the state of conviction.

There're some type errors in the dataset so I need to fix them first. And I need to deal with a lot of NAs exist. For example, when an "age" is not available, I replace it with the column mean. When an Amount of compensation misses, I use "0" to replace it because 0 is the mode. If there're too many NAs in one row, I delete it for simplicity.

What's more, I remove some variables that are useless for my analysis such as First Name and Last Name.

3. Some numbers and likelihood of factors

3.1 Some important numbers in general:

Number of incarcerated (not 0 time) exonerees.	1720
Number of incarcerated exonerees serving <=1 year	184
Number of incarcerated exonerees serving >1 year	1536

We can see that about 90% people who were wrongly convicted of crimes incarcerated. Among those people, about 90% of them serving more than one year.

3.2 Some numbers of exonerees by race and gender:

Number of total exonerees by race/gender				Number of total incarcerated by race/gender			
Asian	13	Female	180	Asian	11	Female	132
Black	921	Male	1720	Black	865	Male	1588
Caucasian	723			Caucasian	641		
Hispanic	222			Hispanic	185		
Native American	12			Native American	10		
Other	9			Other	8		
Average number of years lost by race/gender							
Asian	7.24	Female	4.48				
Black	10.63	Male	9.53				
Caucasian	7.78						
Hispanic	6.79						
Native American	10.225						
Other	6.62						

For both exonerees and incarcerated cases, Black is the biggest race. Next are Caucasian and Hispanic. Asian, native American and other are fewer than two dozen of people. We should focus on the three primary races. By gender, more than 90% of them are male.

About average number of years lost, the black has the highest while Hispanic has the lowest. The male has almost double years lost of the female.

3.3 Some numbers and percentages of exonerees by other factors:

	civil		State		
	Filing	Prevailing	Filing	Prevailing	total
CIU/yes	60/25.5%	28/11.9%	82/34.9%	64/27.2%	235
No	730/43.9%	416/25%	696/41.8%	505/30.3%	1664
guilty pleas/yes	100/24.9%	71/17.7%	97/22.7%	71/17.7%	401
No	690/46%	373/24.9%	681/45.43%	498/33.2%	1499
IO/yes	0/0%	0/0%	0/0%	0/0%	0
No	790/41.6%	444/23.38%	778/41%	569/30%	1899
DNA exonerees/yes	192/55.8%	136/39.5%	248/72%	229/66.5%	344
No	598/38.4%	308/19.8%	530/34%	340/21.8%	1556
death cases/yes	65/56%	34/29.3%	49/42.2%	34/29.3%	116
No	725/40.6%	410/23%	729/40.9%	535/30%	1784

Worst crime/murder	454/57.9%	252/32.2%	360/46%	263/33.6%	783
sexual assault	114/38.5%	66/22.3%	196/66.2%	148/50%	296
drugs	26/12.2%	14/6.5%	22/10.3%	13/6.1%	213
child sexual abuse	72/32.1%	41/18.3%	90/40.1%	61/27.2%	224
robbery	27/28.1%	12/12.5%	45/46.9%	31/32.3%	96
other	97/33.7%	59/20.5%	92/31.9%	53/18.4%	288
Average	38.35%	21.78%	41.10%	30.24%	

This form tells us that:
When the exoneration was not the result of work by a CIU, the likelihood of filing/prevailing on state statutory/civil rights claims is larger.
If one pled his/her guilty, the likelihood is much less than the average.
If the sentence was the death penalty, the likelihood is larger.
If this is a DNA case, the likelihood is much larger than the average.
Among the worst crimes, murder has the largest likelihood and drugs has the lowest likelihood.

4. Relationship between factors and compensation outcome

4.1 Race

Race		state			civil	
	filing	prevailing	amount	filing	prevaling	amount
p-value of ChiSq	<.0001	<.0001	<.0001	0.0016	0.0389	0.1054
Asian	36.36%	27.27	27168	27.27	27.27	
Black	47.01%	35.88	2867	46.59	26.29	
Caucasian	28.99%	16.28	14093	36.28	19.53	
Hispanic	37.04%	31.22	30298.5	42.33	24.34	
Native Am	27.27%	9.09	280	45.45	9.09	
Other	20%	20	12500	0	0	
Effect(Black vs Hispanic)		State			Civil	
Odds Ratio Estimates	filing	prevailing	State amount	filing	prevailing	Civil amount
p-value of Chi-Sq	0.0145	0.0015	0.17	0.2953	0.5865	0.43
Point Estimate	1.5	1.68				
95% Wald Confidence Limits	1.1~2.1	1.22~2.32				
Effect(Black vs Caucasian)		State			Civil	
Odds Ratio Estimates	filing	prevailing	State amount	filing	prevailing	Civil amount
p-value of Chi-Sq	<0.0001	<0.0001	<0.0001	0.0001	0.0032	0.0005

Point Estimate	2.172	2.436	2.7	1.532	1.469	1.586
95% Wald Confidence Limits	1.736~2.719	1.95~3.04	2.09~3.50	1.233~1.904	1.137 1.897	1.222 2.060
Effect(Race*Exonerated)						
p-value of Chi-Sq:	0.035	<.0001	0.0032	0.3845	0.5654	0.2091
Impact on time lost						
Effect(Black vs Hispanic)		State			Civil	
Odds Ratio Estimates	filing	prevailing		filing	prevailing	
p-value of Chi-Sq	<0.0001	<0.0001		0.0001	0.0068	
Point Estimate	1.547	2.436		2.247	1.722	
95% Wald Confidence Limits	0.989~2.419	1.95~3.04		1.467~3.441	1.188~2.496	
Effect(Caucasian vs Hispanic)		State			Civil	
Odds Ratio Estimates	filing	prevailing		filing	prevailing	
p-value of Chi-Sq	<0.0001	0.0661		0.0001	0.0001	
Point Estimate	1.046			1.233	1.309	
95% Wald Confidence Limits	0.650~1.683			0.794~1.915	0.906~1.891	

First, we can see from the contingency table that in general race is associated with the likelihood of state/civil filing/prevailing and the amount of state award because the small p-value of Chi square test. We should reject the null hypothesis which is the row and column variables are independent.

Because of the number of people in the dataset, we should focus on Black, Caucasian and Hispanic. Among the three most important races, Black has the highest likelihood and Caucasian has the lowest.

Conclusion:

1.For black people, the odds of state filing and prevailing are about 1.5 times of the odds for the Hispanic filing and prevailing. The two races have the same level for other cases.

2.On the other hand, for black people, the odds of state filing, prevailing and the state amount are more than 2 times of the odds and amount for the Caucasian. The odds of civil filing/prevailing and civil amount are about 1.5 times.

3.Note that the interaction terms *Race*year of exoneration* are not statistically significant in civil cases. This indicates that there is no evidence that race affects likelihood and amount of civil differently over time. But the relationship has changed in state cases over time.

4.Among people who file and prevail, the Black has 1.5-2.5 times of years lost of the Hispanic. The Caucasian has the same level as the Hispanic.

4.2 Gender

Sex		state			civil	
	filing	prevaling	amount	filing	prevaling	amount
p-value of ChiSq	<.0001	<.0001	0.0526	0.0044	0.0068	0.0083
Odds Ratio Estimates(Female vs Male)	0.43	0.47	0.655	0.62	0.55	0.5
95% Wald Confidence Limits	0.3~0.6	0.33~0.66	0.43~1.00	0.44~0.86	0.36~0.85	0.30~0.84
Effect(Race*Exonerated)						
p-value of Chi-Sq:	<.0001	<.0001	<.0001	0.0079	<.0001	<.0001
Impact on time lost						
Effect(Female vs Male)		State			Civil	
Odds Ratio Estimates	filing	prevailing		filing	prevailing	
p-value of Chi-Sq	<0.0001	0.0003		<0.0001	0.0003	
Point Estimate	0.464	0.551		0.434	0.284	
95% Wald Confidence Limits	0.308 0.698	0.220~1.379		0.311~0.607	0.142~0.566	

Conclusion:

1.Gender and the likelihood of filing/prevailing a state/civil rights claim are strongly associated.
2.For a female, the odds of filing /prevailing are about 0.4~0.6 times of the odds for a male filing/prevailing.

3.Gender is significant in the amount of civil, the average amount of female is about the half of male. But gender isn't significant in the amount of state since the p-value of Chi-Sq test is larger than 0.05. The amount of state and gender are independent.

4.Note that the interaction terms *Sex*year of exoneration* are all statistically significant. This indicates the relationship has changed over time in both state and civil cases.

5.The female has 40%-55% years lost of the male that are in state filing/prevailing and civil filing cases. And the female has only 28% years lost of the male in civil prevailing cases.

4.3 Blue/Red region & Geographic area

Blue/Red region		state			civil	
	filing	prevailing	amount	filing	prevaling	amount
p-value of ChiSq	<.0001	0.037	<.0001	<.0001	<.0001	<.0001
Odds Ratio Estimates (Red vs Blue)	0.37	0.81	0.54	0.31	0.28	0.28
95% Wald Confidence Limits	0.3~0.46	0.66~0.98	0.43~0.68	0.25~0.38	0.22~0.36	0.22~0.37
		state			civil	
Geographic area	filing	prevailing		filing	prevailing	
p-value of ChiSq	<.0001	<.0001		<.0001	<.0001	
Northeast	55.66	35.85		54.72	35.53	
Midwest	46.45	34.52		56.35	30.96	
South	29.83	23.17		22.17	9.67	
West	24.52	14.56		47.51	26.82	
Effect(northeast vs south)		State			Civil	
Odds Ratio Estimates	filing	prevailing	State amount	filing	prevailing	Civil amount
p-value of Chi-Sq	<0.0001	0.1	0.0166	<0.0001	<0.0001	<.0001
Point Estimate	2.952		1.438	4.243	5.151	5.337
95% Wald Confidence Limits	2.227~3.914		1.068~1.936	3.165~5.687	3.611~7.346	3.681~7.739

Conclusion:

1.Blue/Red region and likelihood/amount are all strongly associated.

2.The odds of red filing/prevailing are only about 30% of the odds of blue except state prevailing (80%). And about the average amount, the red only has 50% and 30% of blue relatively.

3.Geographic area and the likelihood of filing/prevailing a state/civil rights claim are strongly associated.

4.Northeast has the highest likelihood in state filing/prevailing and civil prevailing and South has the lowest likelihood in civil filing/prevailing. In general, Northeast and Midwest have much higher rates than South and West.

5.Specifically, in northeast, the odds of state filing are about 3 times of the odds in south. For the likelihood of civil filing/prevailing, the odds of northeast are about 4-5 times of the odds in south. For other geographic areas, the interpretations are similar.

4.4 Other factors

		state		civil
1.CIU	filing	prevailing	filing	prevailing
p-value of ChiSq	0.26	<.0001	<.0001	<.0001
Odds Ratio Estimates(0 vs 1)		0.42	2.33	2.47
95% Wald Confidence Limits		0.32~0.55	1.7~3.2	1.64~3.74
2.DNA	filing	prevailing	filing	prevailing
p-value of ChiSq	<.0001	<.0001	<.0001	<.0001
Odds Ratio Estimates(0 vs 1)	0.174	0.39	0.35	0.3
95% Wald Confidence Limits	0.12~0.24	0.30~0.52	0.26~0.48	0.22~0.40
3.Death	filing	prevailing	filing	prevailing
p-value of ChiSq	0.35	0.09	0.0012	0.1
Odds Ratio Estimates(0 vs 1)			0.53	
95% Wald Confidence Limits			0.36~0.78	
4.guilty plea	filing	prevailing	filing	prevailing
p-value of ChiSq	<.0001	0.68	<.0001	0.005
Odds Ratio Estimates(0 vs 1)	2.43		2.72	1.51
95% Wald Confidence Limits	1.87~3.16		2.1~3.53	1.13~2.03

5.worst crime		state		civil	
	Filing(%)	Prevailing(%)	Filing(%)	Prevailing(%)	Amount($)
p-value of ChiSq	<.0001	<.0001	<.0001	<.0001	<.0001
murder	46.04%	33.63	57.93	32.1	$119,710
sexual assault	46.55%	37.07	26.72	10.34	$44,578
drugs	7.18	5.13	8.72	4.1	$9,208
child sexual abuse	39.91	26.91	32.29	18.39	$37,935
robbery	50	41.18	32.35	23.53	$63,550
other	30	16.09	31.74	19.57	$61,277
Effect(drugs vs murder)		State		Civil	
Odds Ratio Estimates	filing	prevailing	filing	prevailing	Civil amount
p-value of Chi-Sq	<0.0001	0.0404	<0.0001	<0.0001	<.0001

Point Estimate	0.091	0.715	0.069	0.091	0.072
95% Wald Confidence Limits	0.052~0.159	0.51~ 0.985	0.041~0.116	0.044~0.187	0.031~0.165
Effect		State		Civil	
Odds Ratio Estimates(95 %CI)	filing	prevailing	filing	prevailing	Civil amount
child_ vs sexual abuse	0.763(0.485,1.199)	0.570(0.34, 0.936)	1.307(0.795,2.151)	1.952(0.982,3.881)	1.714(0.840,3.498)
drugs vs sexual abuse	0.089(0.046,0.171)	0.098(0.046,0.208)	0.262(0.137,0.499)	0.371(0.147,0.936)	0.407(0.137,1.213)
murder vs sexual abuse	0.979(0.662,1.448)	0.812(0.532,1.239)	3.775(2.444,5.833)	4.097(2.212,7.586)	3.723(1.972,7.030)
other vs sexual abuse	0.492(0.310,0.780)	0.295(0.174,0.502)	1.275(0.776,2.094)	2.108(1.067,4.164)	1.919(0.925,3.982)
robber vs sexual abuse	1.148(0.534,2.467)	1.180(0.521,2.676)	1.311(0.573,3.001)	2.667(0.988,7.194)	2.553(0.939,6.943)

6. year of exoneration		state			civil	
	filing	prevailing	amount	filing	prevailing	amount
p-value of F test	<.0001	<.0001	<.0001	0.65	0.0005	0.4257
Adj R-Sq	0.0029	0.1	0.0098		0.0069	
Parameter Estimates	0.00367	0.02857	13180		-0.00468	
95% Confidence Limits	0~ 0.00673	0.02436~0.03278	6848.4~19511		-0.07~-0.02	
		state			civil	
7. years lost	filing	prevailing	amount	filing	prevailing	amount
p-value of F test	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
Adj R-Sq	0.1303	0.0446	0.0999	0.1282	0.0468	0.0976
Parameter Estimates	0.0211	0.01795	38355	0.02123	0.01101	109440
95% Confidence Limits	0.01841~0.023	0.01388~0.02202	32689~44021	0.01850~0.02396	0.00857~0.01345	93066~125814

Since IO are all equal to 0 in the dataset, we don't need to explore the relationship.

Conclusion:

1.CIU: CIU and the likelihood of filing/prevailing a state/civil rights claim are strongly associated except the state filing. In state prevailing, when CIU=0, the odds are only 40% of CIU=1. But in civil cases, when CIU=0 the rates are about 2.3~2.5 times of CIU=1.

2.DNA: DNA and the likelihood of filing/prevailing a state/civil rights claim are strongly associated. In most cases, when DNA=0, the odds of filing and prevailing are about 30% of DNA=1. In state filing cases, when DNA=0 the odds are only about 15% times of DNA=1.

3.Death: Death is only associated with the likelihood of civil filing. When Death=0, the odds of filing are about 50% of Death=1.

4.Guilty Plea: Guilty plea and the likelihood of filing/prevailing a state/civil rights claim are strongly associated except the state prevailing. For the likelihood of state/civil filing, when Guilty Plea=0, the odds of filing are about 2.5times of Guilty Plea=1. For civil prevailing the odds are about 1.5 times.

5.Worst crime: worst crime and the likelihood of filing/prevailing a state/civil rights claim are strongly associated. Murder has the highest likelihood and amount of civil award while drugs has the lowest. When worst crime=drugs, the odds of filing and prevailing are less than 0.1 times of worst crime=murder except in state prevailing (0.7times in this case). We can also see the comparison between every crime verse sexual abuse. Robbery has the highest odds ratio in state cases (1.1~1.2 times of sexual abuse) while murder has the highest odds ratio in civil cases (3.7~4.1 times of sexual abuse). In general, “drugs” has the weakest effect on filing, prevailing and amount.

6.Year of exoneration: It’s associated with the likelihood of state filing/prevailing, state amount and civil prevailing. In state cases, when time moves on, the likelihood and amount would increase because the confidence intervals of coefficient of the variable in linear regression are always positive.

7.Years Lost: It’s associated with all the likelihood and amount. The coefficient and its confidence interval of years lost in linear regression are all positive, which implies more years lost means higher likelihood and amount.

The F statistic for the overall model is significant, indicating that the model explain a significant portion of the variation in the data for all state/civil cases.

The adjusted R-square indicates the percentage that years lost accounts for the variation in likelihood and amount. In most cases the percentages are about 10%.

5. Multivariate Regression

I fit generalized linear model when the response is the amount of state/civil award and fit logistic regression when the response is the likelihood of state/civil filing/prevailing.

5.1 Amount of state/civil award:

I used forward stepwise method to choose these variables.

1.Amount of state award					
Variable	Parameter Estimate	p-value of F test		Total Variation Accounted For	
Intercept	-163553	<.0001	Semipartial Eta-Square	Omega-Square	90% Confidence Limits
CIU	458375	<.0001	0.0126	0.0121	0.0051~0.0233
Guilty Plea	-172154	0.0032	0.0292	0.0287	0.0170 ~.0441
State Claim Made	807928	<.0001	0.107	0.1065	0.0843~0.1312
0Time	-944816	<.0001	0.0264	0.026	0.0149~0.0408
Denied	-855150	<.0001	0.0514	0.0509	0.0351~0.0701
Pending	-246056	0.005	0.0013	0.0008	0.0000~0.0059
Amount civil	0.024929	0.0022	0.0136	0.0132	0.0057~0.0247
Years Lost	23088	<.0001	0.0248	0.0243	0.0137~0.0388
p-value of F test	<0.0001		Adj R-Sq	0.266377	
2.Amount of civil award					
Variable	Parameter Estimate	p-value of F test		Total Variation Accounted For	
Intercept	-613138	<.0001	Semipartial Eta-Square	Omega-Square	90% Confidence Limits
CIU	256633	0.26	0.0005	0.0001	0.0000~0.0040
DNA_only	510381	<.0001	0.0445	0.0441	0.0293~0.0621
MWID	-482545	<.0001	0.0043	0.0039	0.0006~0.0113
ILD	212959	0.21	0.0006	0.0002	0.0000~0.0044
State Claim Made	243147	<.0001	0.0241	0.0237	0.0131~0.0379
Denied	-468041	0.005	0.0103	0.0099	0.0036~0.0202
Amount	0.21413	<.0001	0.0181	0.0177	0.0088~0.0304
civil Filed	-356171	<.0001	0.0947	0.0943	0.0731~0.1180
No Time	333366	<.0001	0.0609	0.0605	0.0432~0.0808
civil award	3751719	0.0039	0.0033	0.0029	0.0002~0.0097
Years Lost	66007	<.0001	0.0246	0.0242	0.0135~0.0386
p-value of F test	<0.0001		Adj R-Sq	0.266377	

The generalized linear models are:

State amount=-163553+4583375(CIU)-172154(Guilty Plea) +807928(State Claim Made)-944816(0 Time)-855150(Denied)-246056(Pending)+0.025(Amount civil) +23088(Years lost)

Civil amount=-613138+256633(CIU)+510381(DNA only)-482545(MWID)+212959(ILD)+243147(State Claim Made)-468041(Denied)+0.21(Amount state)-356171(Non Statutory Case Filed) +333366(No Time) +3751719(Civil Award) +66007(Years Lost)

The estimated effect sizes tell us the effect of the State Claim Made is greater than the effect of others. “Noncentrality Parameter” and “Partial Variation Accounted For” also tell the same story.

The model interprets that holding other variables at a fixed value, when years lost increases one unit, the amount of receiving state award would increase $23088. The interpretation of other variables is the same. Among all the factors, "State Claim Made" has the biggest positive effect and "Denied" has the biggest negative effect.

Response	Root MSE	R-square	Proportion of Variation Accounted for
Amount of state award	865734.8	0.266377	(0.23,0.29)
log1p (amount of state award)	1.404291	0.941116	(0.94,0.94)
Amount of civil award	2301741	0.378634	(0.35,0.40)
log1p (amount of civil award)	2.09264	0.8751	(0.87,0.88)

To judge the accuracy of predictions, I compute Root MSE value. Because of the large number of the amounts, I used function "log1p=log(x+1)" to make a transformation on the different responses. The output shows that the Root MSE is about 1.5~2.1, which implies the prediction is fairly good for log1p(amount). And the proportion of variation accounted for is 94% for the amount of state award and 88% for the amount of civil award.

5.2 likelihood of state/civil prevailing:

1.Likelihood of State prevailing	Parameter Estimates	Pr > ChiSq	Odds Ratio Estimates	
Intercept	-5.5276	<.0001	Point Estimate	95% Wald Confidence Limits
Age	-0.0261	0.013	0.974	0.954~0.995
DNA only	1.4962	<.0001	4.465	2.506~7.954
State Claim Made	6.9088	<.0001	>999.999	246.408~>999.999
Premature	-1.7592	0.0004	0.172	0.065~0.459
Goodness-of-Fit Statistics	Value/DF	Pr > ChiSq		
Deviance	0.5158	1		
Pearson	0.524	1		
2.Likelihood of civil prevailing	Parameter Estimates	Pr > ChiSq	Odds Ratio Estimates	
Intercept	-6.0615	<.0001	Point Estimate	95% Wald Confidence Limits
CIU	-0.7609	0.0185	0.467	0.248~ 0.880
DNA only	0.7581	0.0102	2.134	1.197~3.805
P_FA	0.4888	0.0535	1.63	0.993~2.678
State Claim Made	0.5968	0.0187	1.816	1.104~2.987
State Award	-0.4911	0.0046	0.612	0.436~0.860
Non Statutory Case File	7.4637	<.0001	>999.999	418.599~>999.999
Dismissed or verdict	-5.8213	<.0001	0.003	<0.001~0.012
Years Lost	-0.0575	<.0001	0.944	0.921~0.968
Goodness-of-Fit Statistics	Value/DF	Pr > ChiSq		
Deviance	0.4968	1		

These are the results of the forward selection process in logistic regression.

The models are:

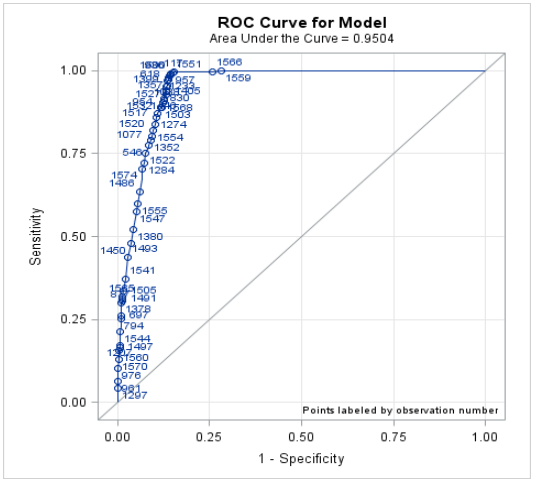
logit(State prevailing)=-5.53-0.026(Age)+1.5(DNA)+6.9(State claim Made)-1.76(Premature)

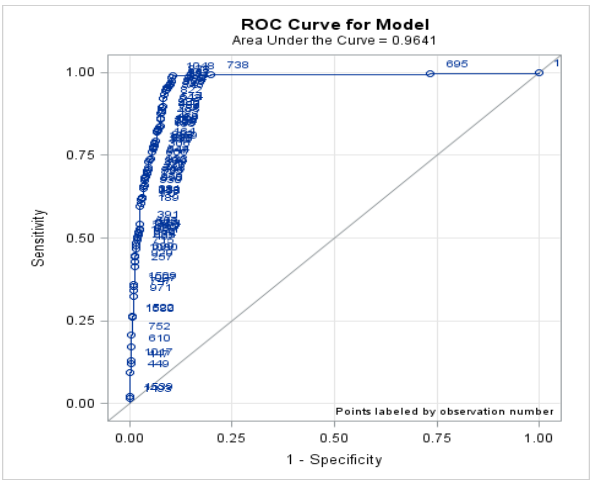
logit(Civil prevailing)=-6-0.76(CIU)+0.76(DNA)+0.49(P_FA)+0.6(State Claim Made)-0.5(State Award)+7.46(Non Statutory Case File)-5.82(Dismissed or verdict)-0.06(Years Lost)

The interpretation of fitted model is that, for instance, holding other variables at a fixed value, the odds of getting state award for people who are in the DNA cases (DNA only = 1) over the odds of getting state award for people who are not (DNA only = 0) is exp(1.5)=4.48. In terms of percent change, we can say that the odds for people who are in DNA cases are 348% higher than the odds for people are not in the cases. The interpretation of other factors is similar.

Odds Ratio Estimates show that “State Claim Made” has the strongest effect on the likelihood of receiving state award, while “Premature” has the lowest effect.

For both goodness of fit statistics, the chi-squares are low relative to the degrees of freedom, and the p-values are high. There is no evidence to reject the null hypothesis, which is that the fitted model is correct.

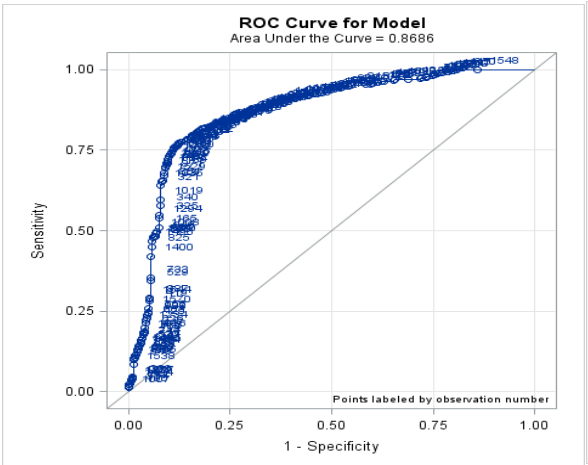


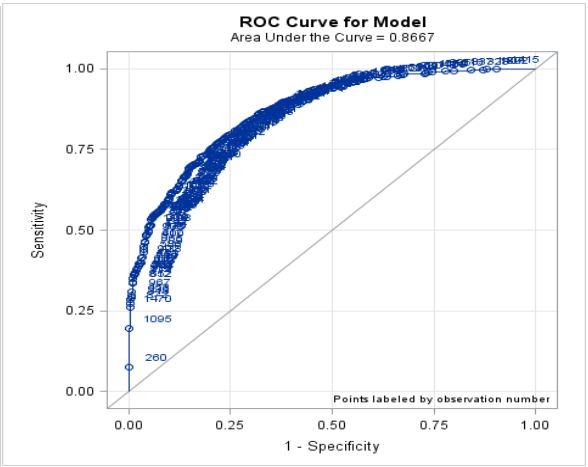


The ROC curves show that the predictions are pretty good. The accuracy of predictions is about 95%.

5.3 likelihood of state/civil filing:

1.likelihood of state filing	Parameter Estimates	Pr > ChiSq	Odds Ratio Estimates	
Intercept	-1.9924	<.0001	Point Estimate	95% Wald Confidence Limits
Guilty Plea	-0.4863	0.0083	0.615	0.429~0.882
DNA only	1.0724	<.0001	2.922	1.922~4.443
MWID	0.645	0.0003	1.906	1.346~2.700
F_MFE	-0.3486	0.0367	0.706	0.509~0.979
P_FA	0.3529	0.0317	1.423	1.031~1.964
State Award	1.8347	<.0001	6.263	5.003~7.841
No Time	-2.3163	0.0016	0.099	0.023~0.417
Premature	-2.6657	<.0001	0.07	0.037~0.129
Years Lost	0.0434	<.0001	1.044	1.026~1.063
2.likelihood of civil filing	Parameter Estimates	Pr > ChiSq	Odds Ratio Estimates	
Intercept	-2.5162	<.0001	Point Estimate	95% Wald Confidence Limits
Age	-0.0143	0.0472	0.986	0.972~1.000
DNA only	0.9084	<.0001	2.48	1.629~3.776
FC	0.8172	<.0001	2.264	1.560~3.285
P_FA	0.7791	<.0001	2.179	1.600~2.968
OM	1.4301	<.0001	4.179	3.135~5.571
State Claim Made	0.3822	0.0075	1.465	1.108~1.939
Dismissed or verdict	5.5709	<.0001	262.674	63.269~>999.999
Years Lost	0.0578	<.0001	1.06	1.041~1.078





These are the results of the forward selection process in logistic regression.

The models are:

logit(State filing)=-2-0.49(Guilty Plea)+1(DNA)+0.64(MWID)-0.35(F_MFE)+0.34(P_FA)+1.83(State Award)-2.3(No time)-2.67(Premature)+0.04(Year lost)

logit(Civil filing)=-2.5-0.01(Age)+0.9(DNA)+0.8(FC)+0.78(P_FA)+1.43(OM)+0.38 (State Claim Made)-5.57(Dismissed or verdict)+0.06(Years Lost)

The interpretation of the models is similar as above.

The ROC curves show that the predictions are pretty good. The accuracy is about 87%.

5.4 Random Forest Prediction

Random Forest	
likelihood of State Award	likelihood of Civil Award
50% training 50% validation	80% training 20% testing
mtry=18 ntree=500	mtry=12 ntree=500
AUC score=0.99	AUC score =0.99

I apply Random Forest model to predict whether a person will receive state/civil award. I randomly subsample training data and testing data and tune parameters such as “mtry” and “ntree” in the model. Finally I use AUC score as a goodness of prediction. The result is that the accuracy of prediction is almost 100%.

6. conclusion

This report explored the specific numbers, percentages and relationship between important factors and the compensation outcome of the wrongly convicted of crimes. If we have more focus on the effect of significant factors such as Race, geographic area, DNA cases and so on, we may reduce these tragic errors in the future. After fitting multivariate regression and making some predictions, I find out that it’s simpler to interpret and predict the likelihood of the compensation by using logistic models. About the amount of compensation, we need plenty of variables and it’s difficult to explain and predict because of lots of randomness by using linear regression models. But if we make a log transformation on the responses, the RMSE of predictions in “glm” model is small, which implies it’s useful in this case.