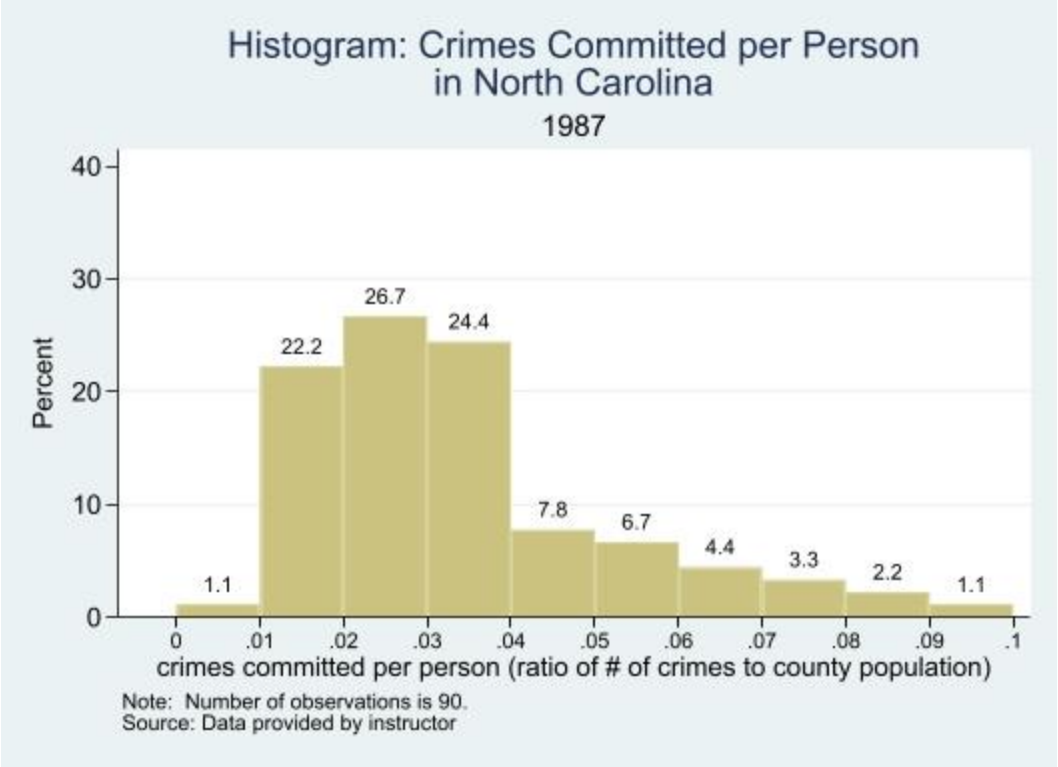
Authors: Dinmukhamed Alchinbayev, Qinlin Wang, Sizhe Sun

Table 1. Summary statistics: North Carolina Crime Data in 1987

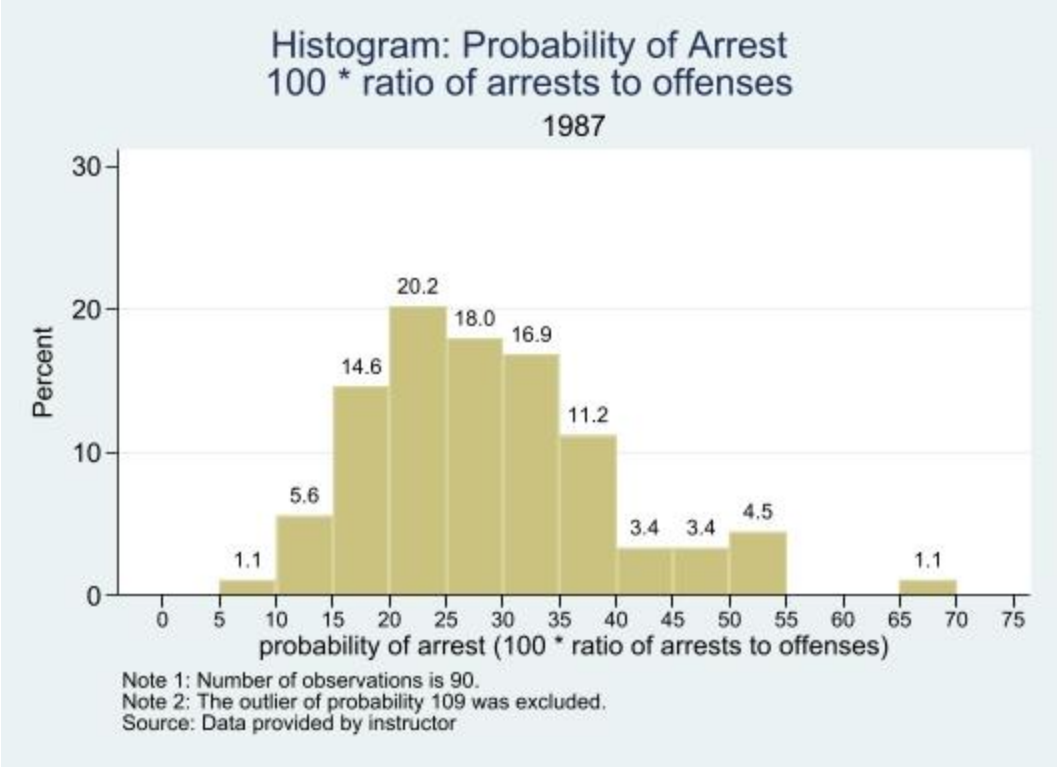
	count	mean	sd	skewness	min	p5	p25	p50	p75	p95	max
crimes committed per person (ratio of # of crimes to county population)	90	0.03	0.02	1.28	0.01	0.01	0.02	0.03	0.04	0.07	0.10
'probability' of arrest (100 * ratio of arrests to offenses)	90	29.52	13.77	2.53	9.28	14.91	20.42	27.15	34.54	52.27	109.09
probability of conviction (100 * ratio of convictions to arrests)	90	55.09	35.42	2.04	6.84	20.78	34.30	45.17	58.89	135.81	212.12
'probability' of prison sentence (100 * ratio of prison sentences to convictions	90	41.06	8.07	-0.45	15.00	25.00	36.36	42.22	45.83	54.29	60.00
average sentence, days	90	9.69	2.83	1.00	5.38	6.18	7.36	9.11	11.51	14.62	20.70
police per capita	90	0.00	0.00	4.98	0.00	0.00	0.00	0.00	0.00	0.00	0.01
thousands of people per sq. mile	90	1.44	1.52	2.65	0.20	0.35	0.55	0.98	1.57	5.12	8.83
tax revenue per capita	90	38.16	13.11	3.29	25.69	27.28	30.70	34.92	41.07	61.15	119.76
=1 if in western region of N.C.	90	0.23	0.43	1.26	0.00	0.00	0.00	0.00	0.00	1.00	1.00
=1 if in central region of N.C.	90	0.38	0.49	0.50	0.00	0.00	0.00	0.00	1.00	1.00	1.00
=1 if in east region of N.C.	90	0.39	0.49	0.46	0.00	0.00	0.00	0.00	1.00	1.00	1.00
=1 if in SMSA	90	0.09	0.29	2.89	0.00	0.00	0.00	0.00	0.00	1.00	1.00
perc. minority, 1980	90	25.71	16.98	0.37	1.28	2.40	10.00	24.85	38.22	58.08	64.35
percent young male	90	0.08	0.02	4.56	0.06	0.07	0.07	0.08	0.08	0.13	0.25
weekly wage, construction industry	90	285.35	47.75	0.61	193.64	218.89	250.73	281.16	315.16	372.16	436.77
wkly wge, transportation, utililty, and communications industry	90	410.91	77.36	0.07	187.62	294.67	374.03	404.78	441.60	548.99	613.23
wkly wge, whlesle, retail trade industries	90	210.92	33.87	1.46	154.21	168.27	190.56	202.99	224.72	277.29	354.68
wkly wge, finance, insurance, real estate industries	90	321.62	54.00	0.82	170.94	254.52	284.59	317.13	342.68	435.11	509.47
wkly wge, service industry	90	275.34	207.40	8.70	133.04	183.15	229.02	253.12	278.11	347.66	2177.07
wkly wge, manufacturing industry	90	336.03	88.23	1.42	157.41	235.05	288.32	321.05	360.21	560.78	646.85
wkly wge, federal employees	90	442.62	59.95	0.13	326.10	338.91	397.33	448.85	478.48	548.49	597.95
wkly wge, state employees	90	357.74	43.29	0.36	258.33	283.90	329.22	358.40	383.72	426.47	499.59
wkly wge, local gov emps	90	312.28	28.13	0.30	239.17	267.08	297.19	307.65	329.16	362.99	388.09
Observations	90										

Source: Data provided by the instructor.

Based on the database which contains data on measures of criminal activity, law enforcement efforts, punishment severity, and several socio-economic variables for 90 counties in North Carolina for the year 1987, we summarized the number of observations, mean, standard deviation, skewness, minimum value, P5, P25, P50, P75, and maximum value for 23 variables by using STATA.



According to the data of crimes committed per person of 90 counties in North Carolina in 1987, we draw the above data distribution histogram. From the histogram, it is clear to see that 75 percent of North Carolina counties have crime rates per person below 4 percent. Only 1.1 percent of the observations (or 1 county) had a crime rate below 0.01 and only 1 county had above 0.09.



According to the data on the probability of arrest in 90 counties in North Carolina in 1987, we draw the above data distribution histogram. From the histogram, it is clear to see that 86.5% of North Carolina counties are above 10 and below 40% of the probability of arrest. Only 1.1 percent of the observations (or 1 county) had a probability of arrest of around 70. Almost 50% of the 90 counties in North Carlina have a probability below 30.

Table 1. Correlation between Crime Rate and some of the potential main explanatory variables

	crmrte	prbarr	prbconv	prbpris	avgsen	polpc
crmrte	1					
prbarr	-.3952849	1				
prbconv	-.3859683	-.0557962	1			
prbpris	.0479934	.0458333	.0110226	1		
avgsen	.0197964	.1786942	.1558523	-.0946809	1	
polpc	.1672764	.4259315	.1718855	.0481343	.4881673	1
N	90					

Source: Data provided by the instructor.

According to the correlation matrix table displaying the correlation between the crime rate and some potential main explanatory variables influencing the crime rate, the probability of arrest (prbarr) and the probability of conviction (prbconv) are relatively significantly correlated with the crime rate (reaching -0.395 and -0.385, respectively). On the other hand, the probability of sentencing (prbpris) and the average sentence (days) (avgsen) is not significantly correlated with the crime rate. It follows that hiring additional security personnel is correlated with the increase in the probability of criminals being arrested and prosecuted. However, the correlation between the crime rate and the two variables of the probability of sentencing and the average sentence period, shows that the effects of punishment severity on the crime rate variable might not be very substantial.

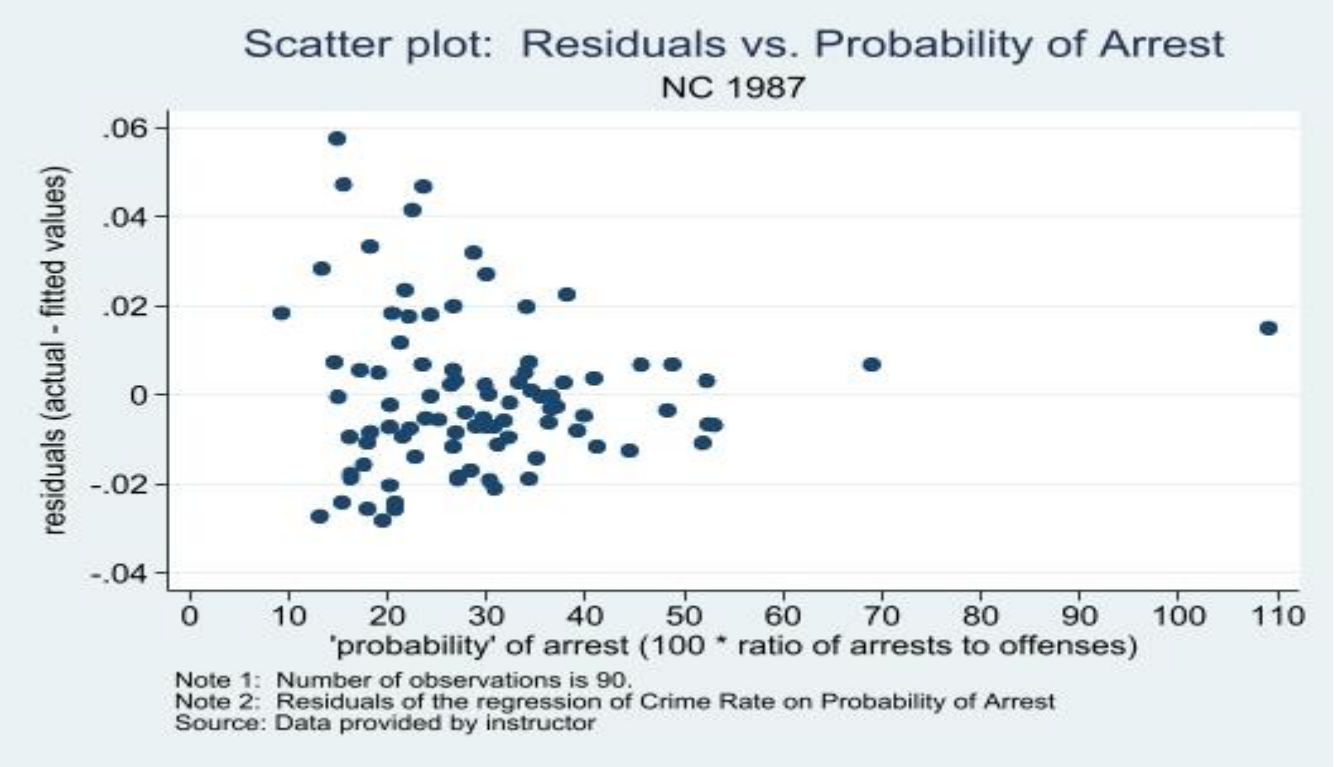
```
. reg crmrte prbarr, r

Linear regression                               Number of obs     =         90
                                                F(1, 88)          =      17.98
                                                Prob > F          =     0.0001
                                                R-squared         =     0.1563
                                                Root MSE          =     .01745
```

crmrte	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
prbarr	-.0005423	.0001279	-4.24	0.000	-.0007964	-.0002881
_cons	.0495201	.0049844	9.94	0.000	.0396147	.0594256

The rise in the probability of arrest by 1 percent, holding everything else constant, decreases the number of crimes to county population by 0.00054.

Based on the table, the p-value is 0.000, which is smaller than the 0.05 level of significance. Given that, prbarr is statistically significant at a 5% level. To identify the practicality of this effect we use the standard deviation of the crimes rate to measure meaningful change. The standard deviation of the crime rate is 0.02 (from the summary statistics above). So, to have a decrease in crimes rate of 1 standard deviation, holding everything else constant, we would need the increase in the probability of arrest by 40 (0.02/0.0005=40). This is almost an 88th percentile of the probability of arrest, or an increase by 3 standard deviations, which is a very big jump (40/13.77=2.9). Thus, the increase in the probability of arrest might not be the most practical tactic to reduce the crime rate.

(4/100 - 5/100) / (-0.0005423) = 18.44
So, the “probability of arrest” should be increased by 18.44% within the respective county in order to achieve the reduction in the crime rate desired by the Sheriff’s Department.
Calculations: Crime rate = 0.0495201 + (-0.0005423)*prbarr + error.
We have the current level of the crime rate of around 5 crimes per 100 citizens if prbarr = 0
Crime rate = 0.0495201 - 0 = 0.05 (rounded). So, if we want to change the 0.05 to 0.04, we need to find the prbarr that will give the difference of 0.01 when multiplied by the coefficient of prbarr.
Prbarr = 0.01/0.0005423=18.439=18.44 (rounded).



Based on this plot we can see that the residuals are heteroskedastic, or in other words, the conditional variance of the residuals is different across the plot. Given that, robust standard errors are the right approach.

Table 1. Regression results for the Crime Rates data from North Carolina in 1987

	(1) crimes committed per person (ratio of # of crimes to county population)	(2) crimes committed per person (ratio of # of crimes to county population)	(3) crimes committed per person (ratio of # of crimes to county population)	(4) crimes committed per person (ratio of # of crimes to county population)
'probability' of arrest (100 * ratio of arrests to offenses)	-0.000542***	-0.000266***	-0.000271***	-0.000383***
	(0.000128)	(0.0000767)	(0.0000784)	(0.0000859)
thousands of people per sq. mile		0.00832***	0.00752***	
		(0.000770)	(0.00129)	
=1 if in SMSA			0.00511	0.0368***
			(0.00804)	(0.00620)
Constant	0.0495***	0.0294***	0.0302***	0.0415***
	(0.00498)	(0.00379)	(0.00407)	(0.00366)
Observations	90	90	90	90
R^2	0.156	0.564	0.566	0.453
Adjusted R^2	0.147	0.554	0.551	0.440
F	17.98	97.24	74.08	36.01
rmse	0.0174	0.0126	0.0127	0.0141

Note 1: Robust standard errors are displayed in parenthesis.
Note 2: The dependent variable is crimes committed per person (ratio of # of crimes to county population).
Significance levels: * p<0.10; ** p<0.05; *** p<0.01
Source: Data provided by the instructor.

Since the effect of the density variable on crime rate is statistically significant in Model 2 and the effect of the probability of arrest decreased when controlled for density we see that density is an omitted variable in Model 1. Thus, Model 1 is biased; the estimated effect of the probability of arrest includes the effect of density on crime rate and, therefore, is overestimated in Model 1. We can also observe the decrease in the standard error of the coefficient for the probability of arrest if we control for density, indicating that including density helps to estimate the "pure" effect of the probability of arrest on crime rates. Given that, Model 2 is a better model specification since its estimated effect of the probability of arrest on crime rate is controlled for density.

Comparing models (2), (3), and (4) with each other, we find that model 2 has the largest adjusted R-squared value and the smallest root mean standard error value. Model #2's adjusted R-squared value equal to 0.554 means that 55.4% of the variation in crime rate can be explained by the regression of crime rate on the probability of arrest and population density. The smallest RMSE value for the regression in model #2 means that using the probability of arrest and population density yields more accurate

predictions than the predictions of models 3 and 4. Additionally, the standard error for the probability of arrest coefficient is the smallest in Model 2, meaning that this model estimates the true effect of this variable on crime rate more precisely.

For the question of why we believe urban went from being non-significant in Model (3) to being statistically significant at the 1% level in Model (4), we consider the two variables "urban" and population density to be highly correlated. Because people yearn to live in urban areas, many people will choose to move to urban which results in an increase in population density. Therefore, it is difficult to estimate the partial effects of urban on the crime rate. Because of that urban is not statistically significant in Model 3 but is significant in Model 4, where we remove the variable of density.

```
. reg crmrte prbarr density west central east, r
note: west omitted because of collinearity.

Linear regression                               Number of obs     =         90
                                                F(4, 85)          =      82.79
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6590
                                                Root MSE          =     .01129
```

crmrte	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
prbarr	-.0002278	.0000771	-2.96	0.004	-.0003811	-.0000746
density	.0087664	.0007218	12.15	0.000	.0073313	.0102014
west	0	(omitted)				
central	.0044778	.0025023	1.79	0.077	-.0004973	.009453
east	.0142202	.0024638	5.77	0.000	.0093215	.0191189
_cons	.0204092	.0031734	6.43	0.000	.0140997	.0267187

When we run the regression of the crmrte on prbarr, density, west, central, and east, the issue that we encountered is the control variable of “west” was omitted because of collinearity. In this case, the category of “West” is chosen as the baseline category.

```
. reg crmrte prbarr density west central, r

Linear regression                               Number of obs     =         90
                                                F(4, 85)          =      82.79
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6590
                                                Root MSE          =     .01129
```

crmrte	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
prbarr	-.0002278	.0000771	-2.96	0.004	-.0003811	-.0000746
density	.0087664	.0007218	12.15	0.000	.0073313	.0102014
west	-.0142202	.0024638	-5.77	0.000	-.0191189	-.0093215
central	-.0097424	.0031962	-3.05	0.003	-.0160973	-.0033875
_cons	.0346294	.0040062	8.64	0.000	.026664	.0425948

In Part#m, we run the regression of crmrte on *prbarr*, *density*, *west*, and *central.* In this case, the control variable of "east" is the baseline category. From the estimated coefficient of central and west, we can know that the crime rate in the central region, holding the probability of arrest and density constant, is lower than that in the eastern region by 0.0097. In the western region of North Carolina, holding the probability of arrest and density constant, the crime rates are lower by 0.0142 compared to the eastern region.

```
. reg crmrte prbarr density east west, r

Linear regression                               Number of obs     =         90
                                                F(4, 85)          =      82.79
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6590
                                                Root MSE          =     .01129
```

crmrte	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
prbarr	-.0002278	.0000771	-2.96	0.004	-.0003811	-.0000746
density	.0087664	.0007218	12.15	0.000	.0073313	.0102014
east	.0097424	.0031962	3.05	0.003	.0033875	.0160973
west	-.0044778	.0025023	-1.79	0.077	-.009453	.0004973
_cons	.024887	.0041493	6.00	0.000	.0166371	.0331369

Comparing the results we obtained in this regression to the results from part (m), we found that the values of estimated dummy coefficients are the same for the control variables of central and east, but the sign of the values are opposite. In Part#n, we run the regression of crmrte on prbarr, density, west, and east. In this case, the control variable of "central" is the baseline category. In Part#m, when the control variable of "east" was chosen as the baseline category, the estimated coefficients of the central variable are negative 0.0097424; In Part#n, when the control variable of "central" was chosen as the baseline category, the estimated coefficients of the "east" are 0.0097424. Based on the estimated coefficients in different parts (l, m, and n), it is easy to see that the cause of the different signs of the estimated dummy coefficients is the difference in baseline categories. Thus, the difference between the crime rates in the eastern and central regions is 0.0097; the crime rate in the east is higher than in the central region by that number.