Homework 2

1. In this question you will work with the dataset **HW2a.dta**, which contains information about arrest records and characteristics of a group of individuals in 1986. All the variables in this dataset have been labeled. We are interested in understanding the determinants of crime and how effective incarceration is.

Using this information answer the following questions:

(a) Create the variable race equals to 1 if the individual is white, and 0 if not (black or hispanic). Show a table providing summary statistics (mean, standard deviation) of narr86, qemp86, inc86, tottime by race.

race	Freq.	Percent	Cum.
0	1,032	37.87	37.87
.1	1,693	62.13	100.00
Total	2,725	100.00	

race	Mean	SD
0	.5784884	1.044891
	2.118314	1.632889
	45.52016	61.20629
	1.452326	6.170497
.1000000	.2982871	.7022276
	2.425281	1.585849
	60.72558	69.11189
	.4647372	3.255763
Total	.4044037	.8590768
	2.309028	1.610428
	54.96705	66.62721
	.8387523	4.607019

(b) Run the following regression:
narr86i = β0 + β1pcnvi + β2avgseni + β3tottimei + β4ptime86i + β5qemp86i + ϵi Interpret the result of each coefficient.

Source	SS	df	MS	Number of obs	=	2,725
				F(5, 2719)	=	24.29
Model	85.9532425	5	17.1906485	Prob > F	=	0.0000
Residual	1924.39391	2,719	.707757967	R-squared	=	0.0428
				Adj R-squared	=	0.0410
Total	2010.34716	2,724	.738012906	Root MSE	=	.84128

narr86	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
pcnv	-.1512246	.040855	-3.70	0.000	-.2313346	-.0711145
avgsen	-.0070487	.0124122	-0.57	0.570	-.031387	.0172897
tottime	.0120953	.0095768	1.26	0.207	-.0066833	.030874
ptime86	-.0392585	.0089166	-4.40	0.000	-.0567425	-.0217745
qemp86	-.1030909	.0103972	-9.92	0.000	-.1234782	-.0827037
_cons	.7060607	.0331524	21.30	0.000	.6410542	.7710671

B0 - shows the predicted number of assets of someone with no prior convictions, no sentence, no incarceration time, no employment time, and not employed in 1986

β1(pcnv) = -0.1512246
For each additional unit increase in proportion of prior conviction, the predicted number of arrests in 1986 changed by β1 units, ceteris paribus.

β2(avgseni) = -0.0070487
For each additional unit increase in average sentence length, the predicted number of arrests in 1986 changes by β2, ceteris paribus.

β3(tottimei) = 0.0120953
For each additional unit of total time incarcerated, the predicted number of arrests in 1986 changes by β3, ceteris paribus.

β4(ptime86i) = -0.0392585
For each additional unit increase in the proportion of time employed in 1986, the predicted number of arrests changes by β4, ceteris paribus.

β5(qemp86i) = -0.1030909
If the person was employed in 1986, the predicted number of arrest in 1986 changes by β5 compared to someone who was not employed, ceteris paribus.

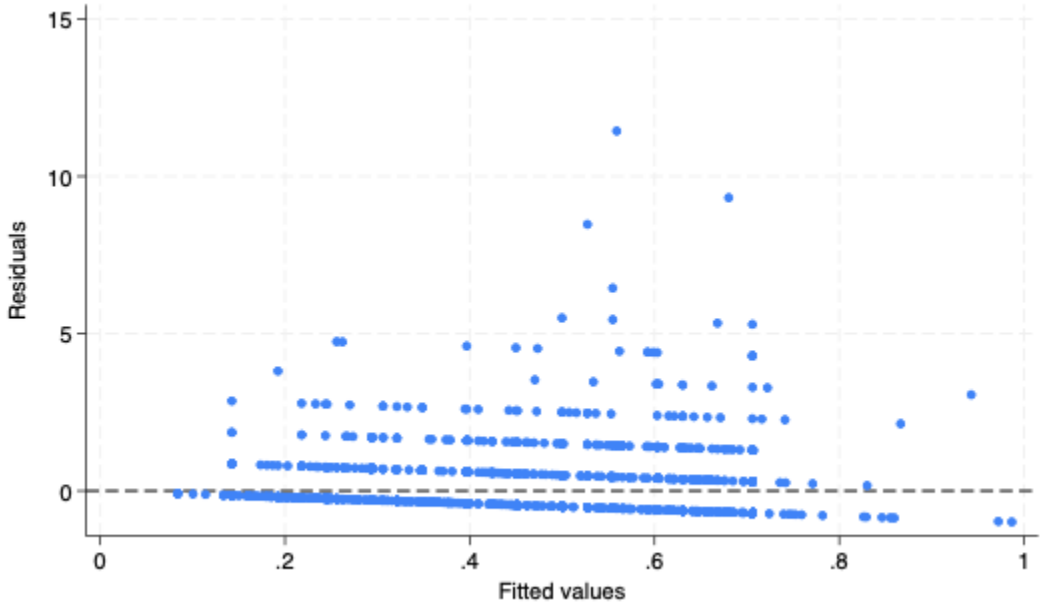
(c) Calculate the standard error for the estimate of avgsen. To do this, you will need to apply the formula seen in class and calculate all the relevant elements separately. (Notice that this is not equivalent to just reporting the value from the regression output).

Avgsen = .00459489

(d) Using your results from part (c) compute a 95% confidence interval for the estimate of avgsen

[-.01605849, .00196117]

(e) Plot the residuals of the regression. Do the errors of the regression appear to be homoskedastic or heteroskedastic? [Hint: you can use the rvfplot command].



The errors seem more heteroskedastic because some spread of the residuals will change as the fitted values increase or decrease.

(f) Create a new variable avgsen2 equals to the average sentence lenght squared. Run the same regression of part (b) including avgsen2. How does the estimate of avgsen change?

```
. regress narr86 pcnv avgsen avgsen2 tottime ptime86 qemp86
```

Source	SS	df	MS	Number of obs	=	2,725
				F(6, 2718)	=	20.71
Model	87.8807329	6	14.6467888	Prob > F	=	0.0000
Residual	1922.46642	2,718	.707309206	R-squared	=	0.0437
				Adj R-squared	=	0.0416
Total	2010.34716	2,724	.738012906	Root MSE	=	.84102

narr86	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
pcnv	-.1537432	.0408706	-3.76	0.000	-.2338837	-.0736027
avgsen	.0146266	.0180657	0.81	0.418	-.0207973	.0500504
avgsen2	-.0005272	.0003194	-1.65	0.099	-.0011534	.000099
tottime	.0065568	.0101447	0.65	0.518	-.0133352	.0264489
ptime86	-.0387868	.0089183	-4.35	0.000	-.0562742	-.0212994
qemp86	-.1025872	.0103984	-9.87	0.000	-.1229768	-.0821976
_cons	.7032527	.0331855	21.19	0.000	.6381813	.7683241

The coefficient of avgsen2 got smaller and reached into the negative.

(g) Run the following regression, employing heteroskedasticity-robust standard errors:

narr86i =α0 + α1pcnvi + α2avgseni + α3avgsen2 i + α4ptime86i + α5qemp86i+ α6inc86i + α7racei + α8tottimei + vi

Is there any change in the estimated coefficient of tottime?

Linear regression

Number of obs	2,725
F(8, 2716)	29.62
Prob > F	0.0000
R-squared	0.0709
Root MSE	.8293

narr86	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
pcnv	-.14091	.0336876	-4.18	0.000	-.2069659	-.074854
avgsen	.0082296	.018623	0.44	0.659	-.028287	.0447462
avgsen2	-.0004479	.0002067	-2.17	0.030	-.0008532	-.0000425
ptime86	-.0408901	.0067811	-6.03	0.000	-.0541868	-.0275933
qemp86	-.0532344	.0144953	-3.67	0.000	-.0816573	-.0248114
inc86	-.0014949	.0002288	-6.53	0.000	-.0019435	-.0010463
race	-2.474556	.3613003	-6.85	0.000	-3.183008	-1.766105
tottime	.0070592	.0135159	0.52	0.602	-.0194433	.0335616
_cons	.8240443	.0495821	16.62	0.000	.7268219	.9212668

Yes, the coefficient went from 0.0120953 to 0.0070592
The decrease of the coefficient may mean that the relationship between time spent and the outcome is weaker.

(h) Name one potential confounder in the specification of part (g). Explain your reasoning

Education is a potential confounder variable because it is associated with the dependent and independent variable. People with less years of education tend to show longer sentences and are more likely to be involved in criminal activity.

2. The dataset HW2b.dta contains information about a job training experiment for a group of men conducted during the years 1976-1977. The objective in this question is to test whether participation in this training program had an effect on the unemployment probability and earnings in 1978.

(a) Create a table showing descriptive statistics (mean and standard deviation) of age, educ, black, married, lre74, lre75, lre78 according to job training status. Do you observe relevant differences between both groups?

-> train = 0

Variable	Obs	Mean	Std. dev.	Min	Max
age	260	25.05385	7.057745	17	55
educ	260	10.08846	1.614325	3	14
black	260	.8269231	.3790434	0	1
married	260	.1538462	.3614971	0	1
lre74	260	.4028257	.8860645	-.583936	3.678089
lre75	260	.2375096	.7819199	-2.480641	3.136885
lre78	260	1.033777	1.111929	-3.106541	3.675883

-> train = 1

Variable	Obs	Mean	Std. dev.	Min	Max
age	185	25.81622	7.155019	17	48
educ	185	10.34595	2.01065	4	16
black	185	.8432432	.3645579	0	1
married	185	.1891892	.3927217	0	1
lre74	185	.4437147	.8883716	-.809299	3.556493
lre75	185	.3327594	.8160968	-2.599059	3.224548
lre78	185	1.279188	1.157459	-1.238599	4.099463

There doesn't seem to be much of a difference between the 2 groups. There is a small increase when train =1, which might suggest that the training program helped decrease unemployment probability and increase earnings in 1978.

(b) Run a simple regression of lre78 on train. What is the estimated effect of participating in job training on real earnings in 1978?

Source	SS	df	MS	Number of obs	=	445
				F(1, 443)	=	5.09
Model	6.50986591	1	6.50986591	Prob > F	=	0.0246
Residual	566.731219	443	1.27930298	R-squared	=	0.0114
				Adj R-squared	=	0.0091
Total	573.241085	444	1.29108353	Root MSE	=	1.1311

lre78	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
train	.2454107	.1087913	2.26	0.025	.0315995	.4592219
_cons	1.033777	.0701455	14.74	0.000	.8959181	1.171637

The people who participated in the job training program had a positive and statistically significant effect on real earnings in 1978 because earnings increased by about 25% with a 95% confidence interval of [0.0315995, 0.4592219]

(c) Calculate the standard error of the estimate of train. (As in Question 1 part (c), this is not the same as reporting the value from the regression output).

standard error of the estimate of train = 2.7206646

(d) Now add as controls to the regression in part (b) the variables re74, re75, educ, age, black, and hisp. Compare your results of the effect of job training with respect to part (b).

Source	SS	df	MS	Number of obs	=	445
				F(7, 437)	=	3.77
Model	32.6286203	7	4.66123147	Prob > F	=	0.0006
Residual	540.612465	437	1.23709946	R-squared	=	0.0569
				Adj R-squared	=	0.0418
Total	573.241085	444	1.29108353	Root MSE	=	1.1122

lre78	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
train	.2321856	.1079703	2.15	0.032	.01998	.4443911
re74	-.001559	.0130975	-0.12	0.905	-.0273009	.0241829
re75	.0281702	.0223654	1.26	0.209	-.015787	.0721274
educ	.0299756	.0299254	1.00	0.317	-.0288401	.0887913
age	.0059764	.0075001	0.80	0.426	-.0087645	.0207172
black	-.6167444	.1977614	-3.12	0.002	-1.005426	-.2280627
hisp	-.0731651	.2637241	-0.28	0.782	-.5914903	.4451601
_cons	1.067114	.4146426	2.57	0.010	.2521724	1.882056

The people who participated in the job training program had a positive and statistically significant effect on real earnings in 1978 because earnings increased by about 23%.

Educated people who participated in the job training program had a negative and statistically significant effect on real earnings in 1978 because earnings decreased by about 61%.

Hispanics who participated in the job training program had a negative and statistically significant effect on real earnings in 1978 because earnings decreased by about 7%.

Hispanics who participated in the job training program had a positive and statistically significant effect on real earnings in 1978 because earnings increased by about 3%.

(e) Create a variable minor equal to 1 if the individual is black or hispanic. Employing this variable incorporate an interaction of minor and train in the regression you run on part (d).

Source	SS	df	MS	Number of obs	=	445
				F(8, 436)	=	3.33
Model	32.9742878	8	4.12178598	Prob > F	=	0.0010
Residual	540.266798	436	1.23914403	R-squared	=	0.0575
				Adj R-squared	=	0.0402
Total	573.241085	444	1.29108353	Root MSE	=	1.1132

lre78	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
train	.0406795	.3783482	0.11	0.914	-.7029336	.7842925
re74	-.0017849	.0131153	-0.14	0.892	-.0275619	.0239922
re75	.0278409	.0223926	1.24	0.214	-.0161699	.0718517
educ	.0304406	.0299631	1.02	0.310	-.0284495	.0893307
age	.0061773	.007516	0.82	0.412	-.0085948	.0209493
black	-.7225105	.2815587	-2.57	0.011	-1.275892	-.1691295
hisp	-.1756718	.327617	-0.54	0.592	-.8195768	.4682332
minor	0	(omitted)				
minor_train	.2083741	.3945257	0.53	0.598	-.5670345	.9837826
_cons	1.156782	.4483696	2.58	0.010	.2755471	2.038016

(f) Repeat the regression of part (d) including controls for unemployment in 1974 and 1975. How does the estimate of train change?

Source	SS	df	MS	Number of obs	=	445
				F(9, 435)	=	2.92
Model	32.6405051	9	3.62672279	Prob > F	=	0.0023
Residual	540.60058	435	1.24275995	R-squared	=	0.0569
				Adj R-squared	=	0.0374
Total	573.241085	444	1.29108353	Root MSE	=	1.1148

lre78	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
train	.2323451	.1085789	2.14	0.033	.0189405	.4457497
re74	-.0009028	.0148956	-0.06	0.952	-.0301791	.0283736
re75	.0280921	.0248423	1.13	0.259	-.0207338	.076918
educ	.0300719	.0302628	0.99	0.321	-.0294076	.0895514
age	.0059079	.0076225	0.78	0.439	-.0090736	.0208895
black	-.6151308	.1990082	-3.09	0.002	-1.006268	-.2239937
hisp	-.069717	.2668389	-0.26	0.794	-.5941708	.4547369
unem74	.0198417	.2029549	0.10	0.922	-.3790525	.4187359
unem75	-.0102496	.1750945	-0.06	0.953	-.354386	.3338868
_cons	1.057004	.4459142	2.37	0.018	.1805896	1.933418

The estimate of train in (d) = 0.23211856, now it is 0.2323451. This means that the estimate of trains slightly increased, adding controls for unemployment in 1974 and 1975. This change is not large enough to indicate a major impact.

(g) Repeat the specification used in part (f), including now unem78 as the dependent variable. Interpret the value of the estimate for train in this case.

Source	SS	df	MS	Number of obs	=	445
				F(9, 435)	=	2.42
Model	4.51942346	9	.502158163	Prob > F	=	0.0109
Residual	90.3030484	435	.207593215	R-squared	=	0.0477
				Adj R-squared	=	0.0280
Total	94.8224719	444	.213564126	Root MSE	=	.45562

unem78	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
train	-.1135081	.0443771	-2.56	0.011	-.2007282	-.026288
re74	-.0022305	.006088	-0.37	0.714	-.0141959	.009735
re75	-.005765	.0101532	-0.57	0.570	-.0257205	.0141906
educ	-.0009564	.0123686	-0.08	0.938	-.0252662	.0233533
age	.0001614	.0031154	0.05	0.959	-.0059617	.0062844
black	.1844998	.0813362	2.27	0.024	.024639	.3443605
hisp	-.0433518	.1090591	-0.40	0.691	-.2577002	.1709965
unem74	.0101425	.0829493	0.12	0.903	-.1528887	.1731737
unem75	.0017219	.0715625	0.02	0.981	-.1389293	.1423732
_cons	.2147717	.1822486	1.18	0.239	-.1434256	.5729691

The coefficient for train(Bo) = -0.1135081. This means that the job training program had a positive effect in reducing unemployment in 1978.