Assignment 3

Econometrics

Elizabeth Goodwin

11/2/2021

Question 1:

This exercise involves data on 299 eruptions of the Old Faithful geyser in Yellowstone National Park. The variable duration is the length of the eruption (in minutes), while the variable waiting is the length of time until the next eruption.

```
use "geyser.dta"
summarize
twoway lfitci waiting duration, plotregion(fcolor(gs15)) ///
ytitle("Waiting Time (Minutes)") xtitle("Duration (Minuites)") ///
legend(off) || scatter waiting duration, mcolor(eltblue)
```

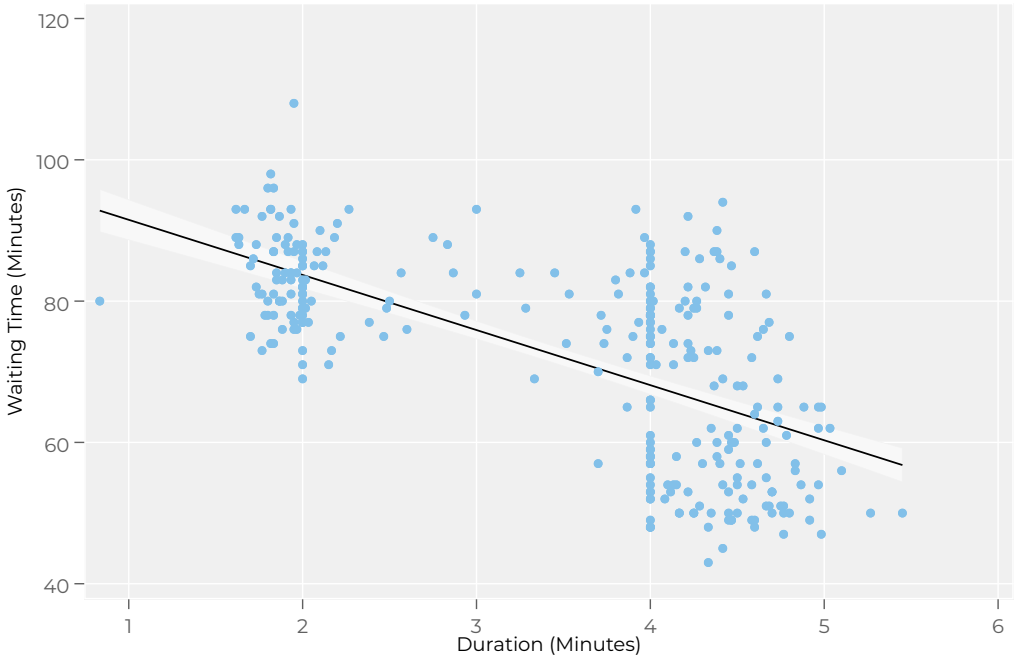

Figure 1: Scatter plot relating duration to waiting time of Geyser eruptions, includes regression line with CIs

There is pretty clear evidence of measurement error in this scatter plot. Primarily, `duration` appears to have many results concentrated with exactly a duration of either 4 or 2. This could be simply due to humans entering data in certain intervals, and rounding to the nearest minute. This doesn't explain why there is nothing at 3, however, as there are two clear clusters around 2 and 4, even on the data that is not exactly two

or four. If on average people people were equally likely to over and underestimate the time due to rounding error, this should not matter for OLS, but it is still pretty clear measurement error. On average, every extra minute of duration relates to a 7.8 minute lower waiting time. This result is highly statistically significant, with a p-value below .001. Running a Breusch-Pagan test clearly shows heteroskedasticity. I adjusted for this with using robust standard errors. The regression table of both regressions is below:

Ran Regression and Breusch-Pagan Test

```
reg waiting duration *Ran Regression
eststo SEs
estat hettest
```

Ran the robust regression and stored table

```
reg waiting duration, vce(robust)
eststo Robust_SEs
esttab using reg.tex, se ar2
```

	(1) SEs	(2) Robust SEs
duration	-7.800***	-7.800***
	(0.537)	(0.454)
Constant	99.31***	99.31***
	(1.957)	(1.394)
N	299	299
adj. R^2	0.414	0.414

Standard errors in parentheses
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Interestingly, the robust standard errors are actually lower than the original standard errors. I think this is because of the unique nature of the way the data is distributed, as σ_i^2 and $(x_i - \bar{x})$ are negatively correlated. In other words, the further away you are from the mean, the lower variance of each observation. The end result of this is that because values further away from the mean count more to the sum of squares. Considering that, Heteroskedasticity of this nature will over count their effect on the error term, since the variance of the observations far away from the mean are lower than the average, and OLS assumes variance is consistent.

You can see this at play in the graph below. As the distance from the mean duration increases, the values of the squared residuals decrease. This is atypical, as in most cases the opposite happens, and the robust SEs are higher than the standard SEs. This problem does not effect robust standard errors, as use the individual squared residuals as the variance instead of assuming constant variance. Thus, in rare cases, where the error trends down as distance from $\bar{x}$ increases, the robust standard error will be lower

```
predict resid, residuals              * Predict Residual Values
gen squared_resid = (resid)^2         * Square Residuals

* Generate Absolute Value from duration from mean
egen mean_duration = mean(duration)
gen dis_duration_mean = abs(duration - mean_duration)

* Graph Results
twoway lfitci squared_resid distance_from_durationmean, plotregion(fcolor(gs15)) ||  ///
scatter squared_resid distance_from_durationmean , mcolor(green)
```

The R^2 for these regressions are both very useful for predicting waiting time. Around 41% of the variance in waiting time can be explained by the duration, which is quite a lot.

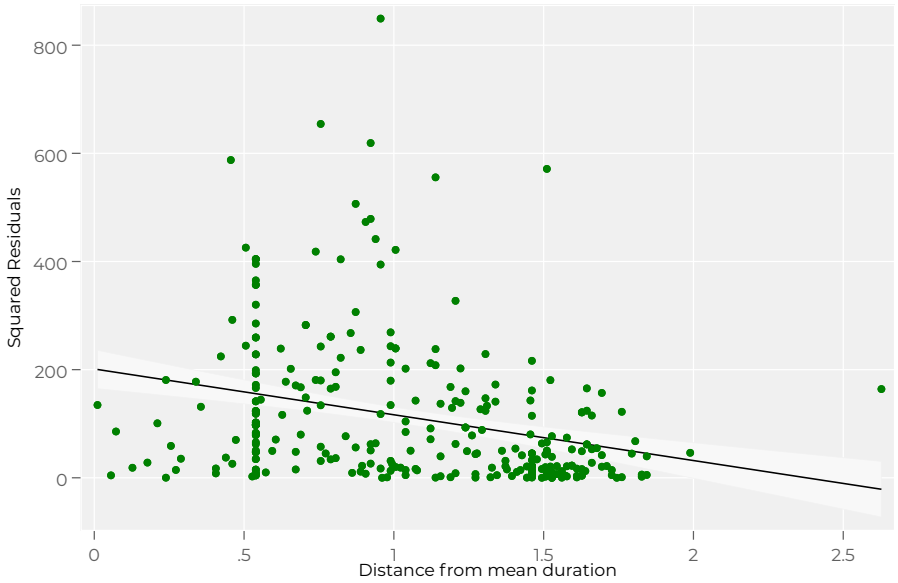

Figure 2: Scatter plot relating the distance from the mean duration and the square residuals, includes regression line with CIs

Question 2: Intergenerational Mobility

First I loaded the dataset and dropped out all below the age of 16:

```
use "linked_1880_1920_males.dta"
drop if age_1 < 16
```

As the data set of father/son occupational income scores are already logged, creating a log-log model of their incomes is just a simple regression between the variables. I also renamed them for easier use. Below are two regressions, the first estimating the elasticity between the father and son's income using their occupation alone, and the second accounting for demographics, region, and industry. Regression table on next page. All regressions ran as robust, as the adjusted scores did not pass a Breusch-Pagan, and wanted to apply the same to both sets.

```
rename ln_occscore_hh Father
rename ln_occscore_child Child
rename ln_adj_occscore_hh Father_Adj
rename ln_adj_occscore_child Child_Adj

eststo: reg Child Father, vce(robust)

eststo: reg Child_Adj Father_Adj, vce(robust)
```

Next I added indicator variables for all the various regions and interaction terms for each. Regression table on the next page. The omitted variable is the New England division.

```
eststo: reg Child Father i.region_1 i.region_1#c.Father, vce(robust)
eststo: reg Child_Adj Father_Adj i.region_1 i.region_1#c.Father_Adj, vce(robust)
```

Many of the results are unfortunately not statistically significant, so it is difficult to tell. In this regression, the important thing is the interaction term, as it tells us the elasticity of each division. The base incomes way be different but $\hat{\beta}_i$ is the elasticity. In the original model, the highest elasticity (lowest mobility) region is the South Atlantic. There are others that are close, such as east south central division, but they are not statistically significant.The lowest elasticity, so the region with the most mobility, is the west south central division, but it is also not remotely statistically significant. So essentially no reasons are statistically significantly more mobile than New England, and only one is statistically less mobile. In the adjusted model, there are more significant values. The south Atlantic and east south central divisions are both far less (higher elasticity) and very statistically significant. No other regions presented statistically significant results from New England.

	(1) Child		(2) Child_Adj		(3) Child		(4) Child_Adj	
Father	0.450***	(0.0358)	0.597***	(0.0333)	0.302**	(0.0970)	0.306**	(0.0933)
middle atlantic division					-0.588	(0.371)	-0.460	(0.347)
east north central div.					-0.424	(0.370)	-0.197	(0.356)
west north central div.					0.00726	(0.440)	-0.245	(0.439)
south atlantic division					-1.013**	(0.358)	-1.470***	(0.316)
east south central div.					-0.928	(0.486)	-1.687***	(0.382)
west south central div.					0.0215	(0.513)	-0.0468	(0.480)
mountain division					-0.428	(0.523)	-0.742	(0.704)
pacific division					-0.672	(0.636)	-0.771	(0.490)
middle atlantic division × Father					0.174	(0.122)		
east north central div. × Father					0.0984	(0.126)		
west north central div. × Father					-0.0669	(0.154)		
south atlantic division × Father					0.291*	(0.119)		
east south central div. × Father					0.251	(0.173)		
west south central div. × Father					-0.121	(0.182)		
mountain division × Father					0.00482	(0.165)		
pacific division × Father					0.189	(0.219)		
middle atlantic division × Father_Adj							0.145	(0.116)
east north central div. × Father_Adj							0.0373	(0.121)
west north central div. × Father_Adj							0.0246	(0.151)
south atlantic division × Father_Adj							0.448***	(0.109)
east south central div. × Father_Adj							0.526***	(0.149)
west south central div. × Father_Adj							-0.143	(0.170)
mountain division × Father_Adj							0.161	(0.238)
pacific division × Father_Adj							0.230	(0.159)
Constant	1.733***	(0.103)	1.305***	(0.0971)	2.295***	(0.293)	2.276***	(0.276)
Observations	1261		1261		1261		1261	
Adjusted R^2	0.142		0.287		0.166		0.345	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

	(1) Child		(2) Child_Adj		(3) Child(SA)		(4) Child_Adj(SA)		(5) Child(ES)		(6) Child_Adj(ES)	
Father	0.0433	(0.147)	0.716***	(0.163)	0.162	(0.231)	0.551**	(0.200)	-0.431	(0.347)	3.640**	(1.176)
white	-0.863*	(0.403)	0.804*	(0.337)	-0.912	(0.659)	0.337	(0.414)	-2.396*	(0.968)	4.321**	(1.630)
white × Father	0.407**	(0.152)			0.434	(0.242)			0.987*	(0.379)		
white × Father_Adj			-0.187	(0.166)			0.112	(0.208)			-2.898*	(1.185)
Constant	2.601***	(0.389)	0.706*	(0.321)	2.210***	(0.621)	0.748	(0.381)	3.765***	(0.873)	-3.501*	(1.594)
Observations	1261		1261		172		172		92		92	
Adjusted R^2	0.150		0.306		0.296		0.497		0.201		0.416	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

When running this regression on race, you get the regression table shown in the page above.

```
eststo: reg Child Father i.race i.race#c.Father, vce(robust)
eststo: reg Child_Adj Father_Adj i.race i.race#c.Father_Adj, vce(robust)
```

The elasticities for white and black people in the non-adjusted model are 0.0433 and 0.4503 respectively. The black elasticity is not at all statistically significant, with $p = .769$. The white elasticity is statistically significant, with $p = .007$. The elasticities for white and black people in the adjusted model are .716 and 0.529 respectfully. The black elasticity is statistically significant, with $p < 0.001$, and the white elasticity not statistically significant, with $p = .262$.

There are very few black people in the sample in general, and most of them do not have an even close to significant sample size. The region with the largest sample of black people is only $n = 16$, with the second largest only being $n = 9$, and most being far smaller. East south central is the closest we have to significant, and the only really even slightly significant sample size, but I included East South even though its low sample size, although I expect non-significant results. I organized these regressions by dropping all but the required regions, and reloading the dataset after running the same regression as before.

For testing nativitiy, we first need to create an indicator variable:

```
gen nat == 0
replace nat = 1 if nativity_1 > 1
drop if nat = 1
eststo: reg Child Father i.region_1 i.region_1#c.Father, vce(robust)
eststo: reg Child_Adj Father_Adj i.region_1 i.region_1#c.Father_Adj, vce(robust)
```

The region with the hightest intergenerationl mobility is south atlantic, and the region with the lowest intergenerational mobility is West South Central. In the adjusted model this changes, and the highest is West South Central and the lowest is Pacific.

It would be difficult to reliably track women across censuses primarily because they are far less likely to have an income. This means that there will be far fewer women with income to track in the first place, meaning a lower sample size, and more importantly the women who do have an income are very unlikely to be representative of the population as a whole. Women who do work in that time period are very likely skewed towards the type of women who do work in that society, which could easily have shared causal factors with whatever you are trying to use said data for.

	(1) Child		(2) Child_Adj	
Father	0.302**	(0.0970)	0.306**	(0.0933)
middle atlantic division	-0.588	(0.371)	-0.460	(0.347)
east north central div.	-0.424	(0.370)	-0.197	(0.356)
west north central div.	0.00726	(0.440)	-0.245	(0.439)
south atlantic division	-1.013**	(0.358)	-1.470***	(0.316)
east south central div.	-0.928	(0.486)	-1.687***	(0.382)
west south central div.	0.0215	(0.513)	-0.0468	(0.480)
mountain division	-0.428	(0.523)	-0.742	(0.704)
pacific division	-0.672	(0.636)	-0.771	(0.490)
middle atlantic division × Father	0.174	(0.122)		
east north central div. × Father	0.0984	(0.126)		
west north central div. × Father	-0.0669	(0.154)		
south atlantic division × Father	0.291*	(0.119)		
east south central div. × Father	0.251	(0.173)		
west south central div. × Father	-0.121	(0.182)		
mountain division × Father	0.00482	(0.165)		
pacific division × Father	0.189	(0.219)		
middle atlantic division × Father_Adj			0.145	(0.116)
east north central div. × Father_Adj			0.0373	(0.121)
west north central div. × Father_Adj			0.0246	(0.151)
south atlantic division × Father_Adj			0.448***	(0.109)
east south central div. × Father_Adj			0.526***	(0.149)
west south central div. × Father_Adj			-0.143	(0.170)
mountain division × Father_Adj			0.161	(0.238)
pacific division × Father_Adj			0.230	(0.159)
Constant	2.295***	(0.293)	2.276***	(0.276)
Observations	1261		1261	
Adjusted R^2	0.166		0.345	

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Question 3

The mean of the robberies is 1.668, the standard deviation is 3.01, and the range is 0 to 87. You can find this with `summarize`

To find the observations containing a street interseciton, do the following

```
tabulate intersection
```

There are 11,102 interactions with a street intersection.

We can do similar for the late hour bars, using `tabulate lh_bar` and counting up the instances. There are 52 observations of a late hour bar.

Using the package `unique`, I did

```
ssc install unique
unique community
```

There are 49 communities represented

Using

```
corr bus_stops rob
```

I found there is a correlation of .4308 between bus stops and robbery counts.

```
gen 1k = .
replace 1k = rob/population
generate standard_pop = (population -   857.2294)/479.5102
generate standard_walkscore = (walkscore - 74.52856)/10.99043
gen com = 0
replace com = 1 if pct_com > .4
```

We chose 40% for our percent commercial, as it seemed large enough to be commercial and on the histogram a large amount were right after 40%.

For our model we decided to use a lasso model. First I standardized all of the coeffients, than ran the `lasso linear` command on it to run the regression

```
lasso linear rob bus_stops lh_bar population intersection  ///
pct_com walkscore, folds(20)
lassocoef, display(coef, standardized)
lassogof
```

Our model predicted with an MSE of 7.092928 and an $R^2 = .2173$. This was actually nearly identical to the same linear model without using lasso, as long as both coefficients were standardized. The coefficients themselves are also quite similar, with `population` decreasing a bit, and most everything else staying very similar. Walkscore is interesting, with it either completely falling out of the model or staying the same depending on when I run it / what seed I use. This is the path of the coefficients.

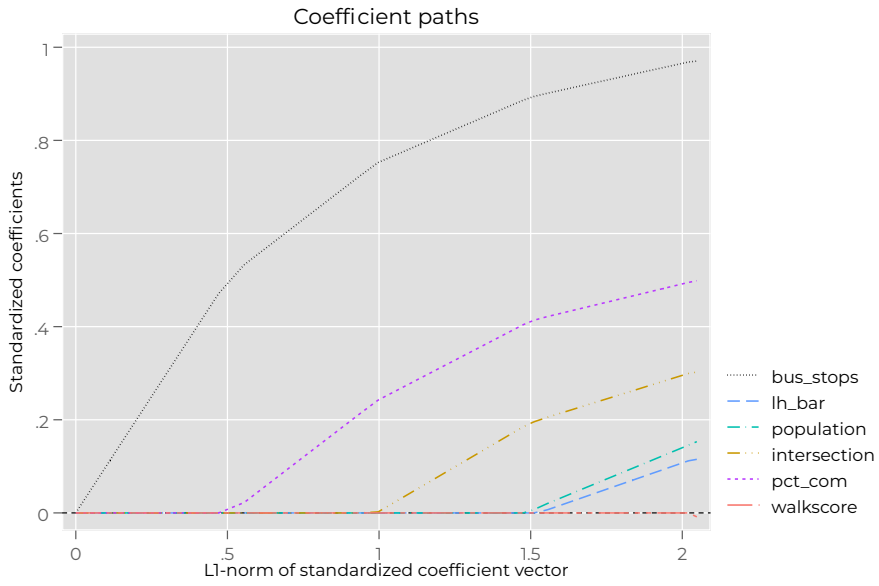

Figure 3: Path of Coefficients

In terms of MSE, The graph is below. λ is very low, so it's not actually that different from the standard regression.

To test the model with differences in commerical uses and population densities, you could break up both commercial and population density values into dummy variables for both. So we have already created this for commercial usage, but you could also break up the standarized population in half to create a "high density" and "low density" categories as well. To do this, we would not want to use a lasso, as inference is the goal for this.

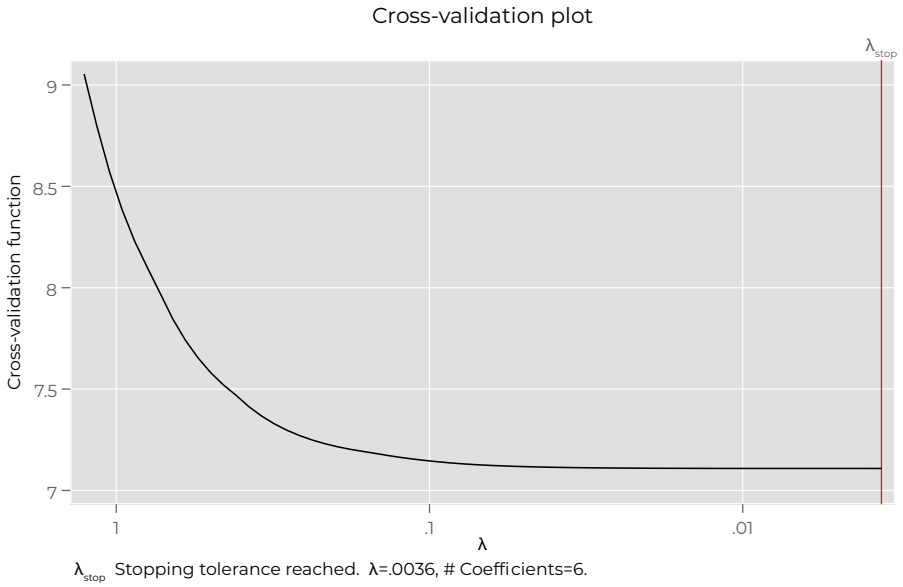

Figure 4: Path of MSE

```
gen com = 0
replace com = 1 if pct_com > .4
gen pop = 0
replace pop = 1 if standard_pop > 0
reg rob com i.pop
```

This is a pretty simple model. It regresses the effect of having high commercial presence (above 40%) on the robberies, and then has an indicator variable for above average population density. To make this more interesting, we could stop working with binary values and use continous ones.

```
reg rob pct_com standard_pop
```

Now, suppose you want to do nonlinear relationships? Well, you could do a few things. You could turn this into an exponential or log model, or you could also do an interaction term between two continous variables. The third coefficent in this means that for every increase in the z-score of population, what is the increase in the rate of change between robberies and commercial density.

```
reg rob pct_com standard_pop c.standard_pop#c.pct_com
```

On the next page you can see the regression tables for all the regressions above. You can make the model less and less linear if you want to, but I overcomplicating it further may not add all that much insight. From this regression, it seems that every increase in percent commercial by itself, there is an additional 3.497 increase in robberies. Adding in a non-interacting population term increases that slightly, and also has a .141 increase in robberies for every increase in the standardized population term. The final model has an additional term, where it suggests that the amount that robberies increases per percent commercial decreases a by .250 for every standard deviation increase in standard pop. This suggests that commerce increases robberies, but as population increases that increase(per percent commerce) gets lower.

	(1) rob	(2) rob	(3) rob	(4) rob	(5) rob
com	2.390***	2.396***			
	(0.0578)	(0.0578)			
pop=1		0.221***			
		(0.0418)			
pct_com			3.497***	3.503***	3.502***
			(0.0744)	(0.0743)	(0.0743)
standard_pop				0.141***	0.179***
				(0.0205)	(0.0233)
standard_pop × pct_com					-0.250***
					(0.0747)
Constant	1.305***	1.208***	1.222***	1.221***	1.221***
	(0.0225)	(0.0291)	(0.0226)	(0.0226)	(0.0226)
Observations	19330	19330	19330	19330	19330
Adjusted R^2	0.081	0.082	0.103	0.105	0.105

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$