Valeria Salinas-Lopez

Fall 2018

Nov 1, 2018

SOC 3305

Lab 5

(Due by 11:59 pm on Nov 8, 2018)

Please save your final work as lastname_lab5.doc (put your name above the title page) and email it to instructor. Please explain your results with your own words. (You must use Stata program to do this lab). Prove your Stata command for each question.

1. **Use student2.dta for this question. Test if mean GPA of a student differs if they belong to greek (fraternity or sorority) or not. Also, test if medians are same or not. Explain your result and also verify it by graphing it. Join your two graphs (greek or non-greek) in one graph. Report your codes. (20 points).**

   Mean Command: **ttest gpa, by (greek)**
   Result:

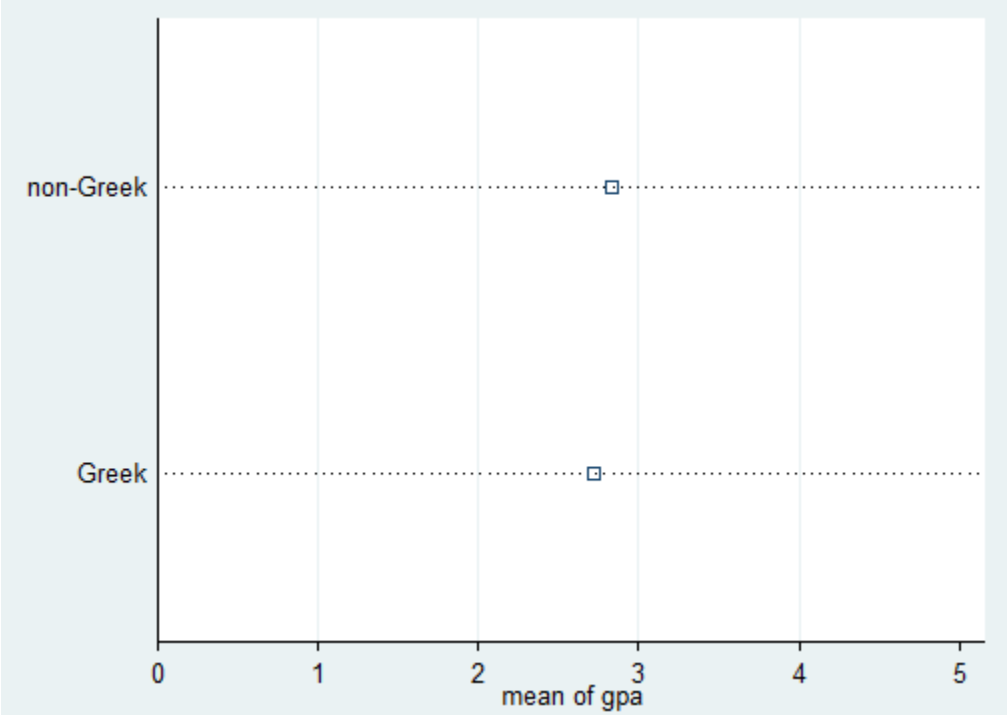
```
. ttest gpa, by(greek)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
non-Gree	173	2.830809	.0345168	.453997	2.762678	2.89894
Greek	45	2.722222	.0706363	.4738426	2.579864	2.86458
combined	218	2.808394	.0310989	.4591705	2.7471	2.869689
diff		.108587	.0766599		-.0425102	.2596842

```
    diff = mean(non-Gree) - mean(Greek)                           t =   1.4165
Ho: diff = 0                                     degrees of freedom =      216

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9210         Pr(|T| > |t|) = 0.1581          Pr(T > t) = 0.0790
```

Interpretation: The null hypothesis here states that the mean gpa of non-greeks, and the mean gpa of greeks, are the same. However, if you look at the probability in the middle, the probability is not significant (it is more than 0.05). This means that we cannot reject the null hypothesis, so we have to fail to reject it. There is not enough evidence to say that the means are, in fact, different.

Graph: graph dot (mean) gpa, over(greek) ylabel(0(1)5, grid) marker(1, msymbol(Sh))
Saving the Graph: graph save Graph "C:\Users\vxs170930\Downloads\MeanGPA.gph"



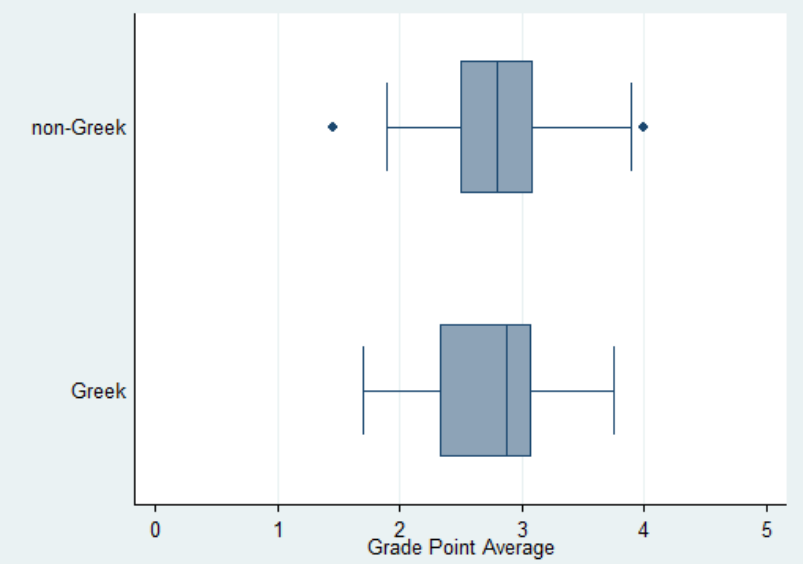
Interpretation: Here you can see that the means are very close to each other. They are not the same, however, there is no considerable difference. Non-greeks have about a 2.8 gpa average, while the greeks have about a 2.7 gpa average. Therefore, the null hypothesis can be rejected, because the means are not the same.

Median Command: ranksum gpa, by(greek)

```
. ranksum gpa, by(greek)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

      greek |      obs    rank sum    expected
------------+---------------------------------
  non-Greek |      173     19298.5     18943.5
      Greek |       45      4572.5      4927.5
------------+---------------------------------
   combined |      218       23871       23871

unadjusted variance    142076.25
adjustment for ties      -231.71
                       ----------
adjusted variance      141844.54

Ho: gpa(greek==non-Greek) = gpa(greek==Greek)
             z =   0.943
    Prob > |z| =   0.3459
```

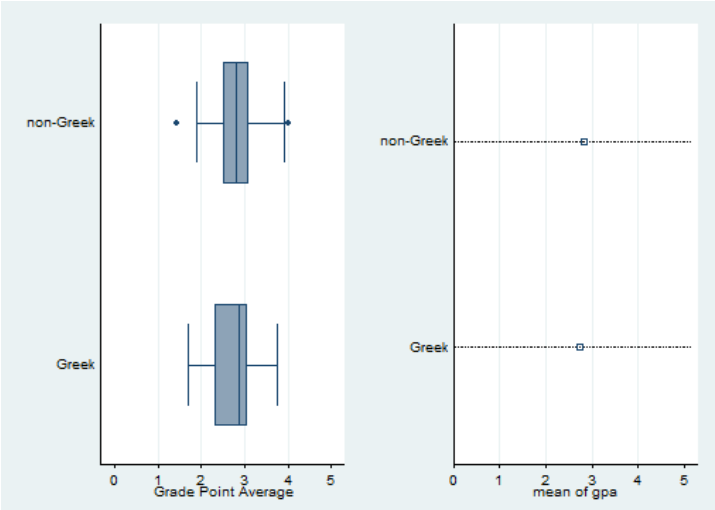
Interpretation: The null hypothesis here states that the median gpa for greeks and non-greeks are equal to each other. The probability here is also not significant since it is greater than 0.05. This means that we do not have enough evidence to reject the null hypothesis that the median gpas are the same, therefore, we must fail to reject the null hypothesis.

Graph: **graph hbox gpa, over(greek) ylabel(0(1)5)**
Saving the Graph: **graph save Graph "C:\Users\vxs170930\Downloads\MedianGPA.gph"**



Interpretation: Here you can see that the median gpa for non-greeks is almost equal to the median gpa for the greeks. Since they are very similar, we had to reject the null hypothesis with the test in stata but once you graph it, you can see that the medians are not equal therefore, the null hypothesis should be rejected.

Combined Graphs: **graph combine C:\Users\vxs170930\Downloads\MedianGPA.gph C:\Users\vxs170930\Downloads\MeanGPA.gp**



2. **Conduct a one-way analysis of variance (ANOVA). Test average aggressiveness between greek and non-greek. Is there a significant difference in mean variance? Is this test valid or not? Why? (20 points)**
   Command: **oneway aggress greek, tabulate**

```
. oneway aggress greek, tabulate
```

Belong to fraternity or sorority	Summary of Aggressive behavior scale		
	Mean	Std. Dev.	Freq.
non-Greek	1.2959184	1.6530263	196
Greek	1.9148936	1.5993176	47
Total	1.4156379	1.657715	243

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Between groups	14.524267	1	14.524267	5.38	0.0212
Within groups	650.496309	241	2.69915481		
Total	665.020576	242	2.74801891		

```
Bartlett's test for equal variances:  chi2(1) =   0.0795  Prob>chi2 = 0.778
```

Interpretation: Since Prob>F is less than 0.05, there is a significant difference in mean variance. However, Prob>chi2 is greater than 0.05, therefore, the test is not valid.

3. **In this Questions use same variables as in Q2 but include gender of a student and also drinking to analysis. Later include interaction of gender, year and drinking, and decide if this interaction term contributes to the model? Does anything change if you define gender as binary and drinking as continuous variables? Also, report margins by year and report marginsplot. (20 points)**
   Command: anova aggress greek gender drink

```
. anova aggress greek gender drink
```

Number of obs =	243		R-squared	=	0.4345
Root MSE =	1.33186		Adj R-squared	=	0.3545

Source	Partial SS	df	MS	F	Prob>F
Model	288.96384	30	9.6321279	5.43	0.0000
greek	4.2131842	1	4.2131842	2.38	0.1248
gender	51.075121	1	51.075121	28.79	0.0000
drink	159.98313	28	5.7136832	3.22	0.0000
Residual	376.05674	212	1.7738525		
Total	665.02058	242	2.7480189		

Commands for Interactions: anova aggress greek gender drink gender#year#drink

```
. anova aggress greek gender drink gender#year#drink
```

Number of obs =	243		R-squared	=	0.7047
Root MSE =	1.3242		Adj R-squared	=	0.3619

Source	Partial SS	df	MS	F	Prob>F
Model	468.62906	130	3.6048389	2.06	0.0001
greek	5.8584813	1	5.8584813	3.34	0.0702
gender	36.90782	1	36.90782	21.05	0.0000
drink	139.05889	28	4.966389	2.83	0.0001
gender#year#drink	179.66522	100	1.7966522	1.02	0.4489
Residual	196.39152	112	1.7534957		
Total	665.02058	242	2.7480189		

Interpretation: I think that the interaction does not add to the model since its probability is greater than 0.05.

Commands for Defining: anova aggress greek i.gender c.drink i.gender#year#c.drink

```
. anova aggress greek i.gender c.drink i.gender#year#c.drink
```

Number of obs =	243	R-squared	=	0.4186	
Root MSE =	1.29098	Adj R-squared	=	0.3935	

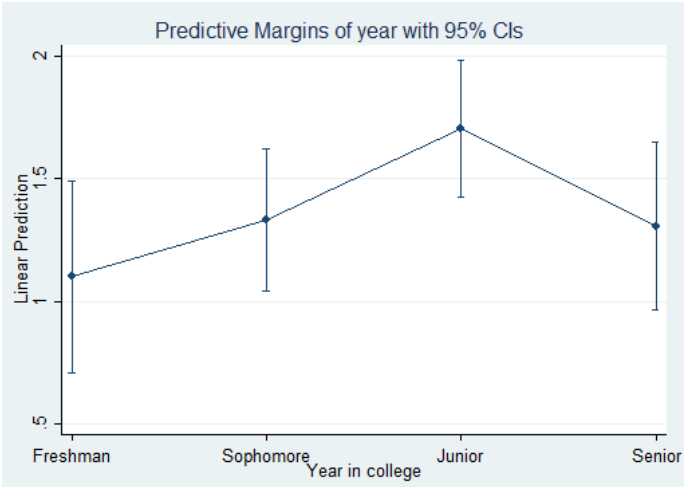
Source	Partial SS	df	MS	F	Prob>F
Model	278.36164	10	27.836164	16.70	0.0000
greek	7.3818781	1	7.3818781	4.43	0.0364
gender	3.6235706	1	3.6235706	2.17	0.1417
drink	101.54211	1	101.54211	60.93	0.0000
gender#year#drink	40.265799	7	5.752257	3.45	0.0016
Residual	386.65894	232	1.6666334		
Total	665.02058	242	2.7480189		

Interpretation: When you define gender as binary and drinking as continuous, the R-squared is suppressed. This means that the independent variables explain less of the dependent variables variance when the variables are defined as binary and continuous.

Commands for Margins: margins year and marginsplot

```
. margins year

Predictive margins                              Number of obs     =        243

Expression   : Linear prediction, predict()
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
year						
Freshman	1.099849	.1993073	5.52	0.000	.7071655	1.492533
Sophomore	1.333668	.1479123	9.02	0.000	1.042245	1.625091
Junior	1.704846	.1428561	11.93	0.000	1.423385	1.986307
Senior	1.305681	.1738141	7.51	0.000	.9632251	1.648137



4. **Define F test, t test, R^2, adj-R^2, Type I and Type II errors, pairwise correlation (pwcorr) and multicollinearity. (20 points)**

t-test: This test's null hypothesis states that the coefficient of a particular independent variable is equal to zero, meaning that the particular x variable has no effect on the dependent variable being tested. If the probability comes out to be 0.05 or less, you reject the null, meaning that the independent variable is not equal to zero and therefore has an effect on the dependent variable.
F-test: This test's null hypothesis states that all of the coefficients of all of the independent variables are equal to zero, meaning that none of them effect the given dependent variable. If the probability comes out to be 0.05 or less, you reject the null, meaning that all of the independent variables are not equal to zero and therefore they all have an effect on the dependent variable.
R2 (r square): This explains the variance of the dependent variable by giving the percentage of the independent variable(s) effect.
Adjusted R2: This is similar to the R2; however, it considers the complexity of the data.
Type 1 error: This is when you reject a true null hypothesis.
Type 2 error: This is when you fail to reject a false null hypothesis.
pwcorr: (pairwise correlation): States the correlation based on all of the observations that are actually recorded, so it gets rid of missing data. With this command you can also select the level of significance you would like.
Multicollinearity: This is when one variable in a multiple regression can be predicted from other variables with greater accuracy.

5. **Use Nations2.dta. See how life expectancy predicts average school years. Write formula and explain your result (F test, t test and significance). Later include GDP, child mortality and adolescence fertility to the model. Report correlation and pairwise correlation (with .05 level of significance) tables and explain. Is there a possibility of committing Type I or Type II error? Transform gdp with natural logarithm and run the regression with it. Come up with the reduced model to predict schooling and report the formula with coefficients. (20 pts)**
   Commands: regress life school

```
. regress life school
```

Source	SS	df	MS	Number of obs	=	188
				F(1, 186)	=	206.34
Model	9846.65406	1	9846.65406	Prob > F	=	0.0000
Residual	8875.86926	186	47.7197272	R-squared	=	0.5259
				Adj R-squared	=	0.5234
Total	18722.5233	187	100.120446	Root MSE	=	6.9079

life	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
school	2.45184	.1706856	14.36	0.000	2.115112	2.788569
_cons	50.35941	1.36924	36.78	0.000	47.65817	53.06065

Formula: $Y_i=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\ldots+\beta_ix_{i+}e_i$

Predicted Life Expectancy= 50.36 + 2.45(years of school)

T-test: Since the probability is 0.000, which is less than 0.05, this means that the independent coefficient of mean of years in school is not equal to zero, and therefore has an effect on the life expectancy.

F-test: Since Prob>F is less than 0.05, the coefficients of all the independent variables are not equal to zero. This means that the independent variables in this model have an effect on life expectancy.

Significance: Since the probability of the mean years of schooling is less than 0.05, this independent variable has a significant effect on life expectancy.

Command: regress life school gdp chldmort adfert

```
. regress life school gdp chldmort adfert
```

Source	SS	df	MS	Number of obs	=	178
				F(4, 173)	=	337.70
Model	15635.6868	4	3908.92171	Prob > F	=	0.0000
Residual	2002.50296	173	11.5751616	R-squared	=	0.8865
				Adj R-squared	=	0.8838
Total	17638.1898	177	99.6507898	Root MSE	=	3.4022

life	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
school	-.1442219	.1438604	-1.00	0.317	-.4281693	.1397256
gdp	.0001338	.000023	5.81	0.000	.0000884	.0001793
chldmort	-.1652258	.0091747	-18.01	0.000	-.1833346	-.147117
adfert	.0041719	.0095542	0.44	0.663	-.0146858	.0230297
_cons	75.70404	1.399551	54.09	0.000	72.94164	78.46643

Command: correlate life school gdp chldmort adfert

```
. correlate life school gdp chldmort adfert
(obs=178)
```

	life	school	gdp	chldmort	adfert
life	1.0000				
school	0.7313	1.0000			
gdp	0.6062	0.5717	1.0000		
chldmort	-0.9294	-0.7724	-0.5160	1.0000	
adfert	-0.7424	-0.6798	-0.5121	0.7888	1.0000

Interpretation:

a. Life expectancy and mean years in school are positively correlated, meaning as one goes up so does the other. I would say that they are kind of strongly correlated since 0.7 is close to 1.
b. Life expectancy and GDP are positively correlated, meaning as one goes up so does the other. These two are sort of strongly correlated since 0.6 relatively close to 1, but it is weaker than that of life expectancy and mean years in school.
c. Life expectancy and child mortality are negatively correlated, meaning as one goes up, the other goes down. I would say they are strongly correlated since -0.9 is close to -1.
d. Life expectancy and adolescent fertility are negatively correlated, meaning as one goes up, the other goes down. They are somewhat strongly correlated since -0.7 is close to -1.

Command: pwcorr life school gdp chldmort adfert , star(.05)

```
. pwcorr life school gdp chldmort adfert , star(.05)
```

	life	school	gdp	chldmort	adfert
life	1.0000				
school	0.7252*	1.0000			
gdp	0.6112*	0.5733*	1.0000		
chldmort	-0.9236*	-0.7727*	-0.5160*	1.0000	
adfert	-0.7318*	-0.6752*	-0.5171*	0.7774*	1.0000

Interpretation: All of the relationships between the variables tested are significant because there is a star next the numbers. None of them are not significant.

The command "sidak" is used to lessen the probability of Type I Error.

Command: pwcorr life school gdp chldmort adfert, sidak sig star(.05)

```
. pwcorr  life school gdp chldmort adfert, sidak sig star(.05)
```

	life	school	gdp	chldmort	adfert
life	1.0000				
school	0.7252*	1.0000			
	0.0000				
gdp	0.6112*	0.5733*	1.0000		
	0.0000	0.0000			
chldmort	-0.9236*	-0.7727*	-0.5160*	1.0000	
	0.0000	0.0000	0.0000		
adfert	-0.7318*	-0.6752*	-0.5171*	0.7774*	1.0000
	0.0000	0.0000	0.0000	0.0000	

Interpretation: All of the correlations seem to be significant since they all have a star next to them. There are none that do not have a star. Mean years in school, GDP, child mortality, and adolescent fertility all significantly impact the life expectancy.

Type I and Type II: The above test does not change after using "sidak," meaning that there is a low chance of a Type I Error. Type I Errors are always more likely than a Type II Error to occur.

Command for Transforming GDP: generate loggdp=log10(gdp)

```
. regress life school gdp chldmort adfert loggdp
```

Source	SS	df	MS	Number of obs	=	178
				F(5, 172)	=	269.62
Model	15642.4196	5	3128.48391	Prob > F	=	0.0000
Residual	1995.77023	172	11.6033153	R-squared	=	0.8868
				Adj R-squared	=	0.8836
Total	17638.1898	177	99.6507898	Root MSE	=	3.4064

life	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
school	-.1847653	.1535547	-1.20	0.231	-.4878595	.118329
gdp	.0001106	.0000382	2.89	0.004	.0000352	.000186
chldmort	-.1616093	.0103403	-15.63	0.000	-.1820195	-.1411991
adfert	.0036516	.0095901	0.38	0.704	-.0152779	.0225811
loggdp	1.00576	1.320352	0.76	0.447	-1.600419	3.611939
_cons	72.34118	4.631775	15.62	0.000	63.19874	81.48362

Commands for Reducing the Model: regress life school, level(99) and margins, at (school=(2 12)) vsquish

```
. regress life school, level(99)
```

Source	SS	df	MS	Number of obs	=	188
				F(1, 186)	=	206.34
Model	9846.65406	1	9846.65406	Prob > F	=	0.0000
Residual	8875.86926	186	47.7197272	R-squared	=	0.5259
				Adj R-squared	=	0.5234
Total	18722.5233	187	100.120446	Root MSE	=	6.9079

life	Coef.	Std. Err.	t	P>|t|	[99% Conf.	Interval]
school	2.45184	.1706856	14.36	0.000	2.007628	2.896053
_cons	50.35941	1.36924	36.78	0.000	46.79594	53.92289

```
. margins, at (school=(2 12)) vsquish

Adjusted predictions                            Number of obs     =        188
Model VCE    : OLS

Expression   : Linear prediction, predict()
1._at        : school          =           2
2._at        : school          =          12
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
_at						
1	55.26309	1.059291	52.17	0.000	53.17332	57.35286
2	79.78149	.9244047	86.31	0.000	77.95783	81.60516