1) Continue practicing interactions. Pick a continuous dependent variable of interest, and two categorical independent variables of interest (these could be continuous variables that you turn into categorical variables, if you wish). Motivate (even in 1 paragraph) a relationship of interest between 1 of the categorical predictors and your dependent variable, and why that relationship might vary according to some other categorical variable. Estimate the additive and interactive models, perform any necessary tests to answer your question, and briefly interpret your results as they relate to your relationship of interest.

*The following analysis is based on GSS respondents who answered the question of "do you have internet access at home" from 2006 - 2014.
Continuous dependent variable: prestg10
Categorical Independent variable 1: educlevel
Categorical Independent variable 1: intraccess

I suspect that occupational prestige score (prestg10) and R's educational level (educlevel) has a positive correlation because people with higher educational attainment can qualify for more jobs with high occupational prestige, such as careers in law, medical science and engineering. At the same time, I believe people who have internet access at home have a higher chance of finding better jobs (increased occupational prestige score) because the internet provides a lot of career guidance. To test the relationship, model 1 (additive) and model 2 (interactive) are shown.

Model 1:

```
. xi: regress prestg10 i.educlevel i.intraccess
i.educlevel       _Ieduclevel_0-3     (naturally coded; _Ieduclevel_0 omitted)
i.intraccess      _Iintracces_0-1     (naturally coded; _Iintracces_0 omitted)
```

Source	SS	df	MS		
Model	330888.385	4	82722.0963	Number of obs	= 6,290
Residual	1236663.93	6,285	196.764348	F(4, 6285)	= 420.41
				Prob > F	= 0.0000
Total	1567552.31	6,289	249.253031	R-squared	= 0.2111
				Adj R-squared	= 0.2106
				Root MSE	= 14.027

prestg10	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
_Ieduclevel_1	.5388606	.9714478	0.55	0.579	-1.365509	2.44323
_Ieduclevel_2	6.433616	.8609409	7.47	0.000	4.745878	8.121354
_Ieduclevel_3	18.61039	.9087487	20.48	0.000	16.82893	20.39185
_Iintracces_1	-3.46955	.4240925	-8.18	0.000	-4.300916	-2.638184
_cons	33.75502	.8631727	39.11	0.000	32.06291	35.44713

```
. test _Ieduclevel_1 _Ieduclevel_2 _Ieduclevel_3

 ( 1)  _Ieduclevel_1 = 0
 ( 2)  _Ieduclevel_2 = 0
 ( 3)  _Ieduclevel_3 = 0

       F(  3,  6285) =  404.75
            Prob > F =    0.0000
```

```
. test _Ieduclevel_1=_Ieduclevel_2=_Ieduclevel_3

 ( 1)  _Ieduclevel_1 - _Ieduclevel_2 = 0
 ( 2)  _Ieduclevel_1 - _Ieduclevel_3 = 0

       F(  2,  6285) =  566.26
            Prob > F =    0.0000
```

Using the additive model, my reference group is R with an education level of below high school and R with internet access at home. Based on the P-value table in the regression, internet access at home, the education level of "below college graduate" and the education level of "college graduate and above" all have a P-value <0.05, therefore we reject the null hypothesis that they do not have an effect on the occupational prestige score at the significance level of 5%, which means that they do have an effect on the occupational prestige score. The education level of "below high school graduate" has a P-value of 0.579 and we are not able to reject the null hypothesis that it does not have an effect on the occupational prestige score at the significance level of 5% but 10%, which means that it does not have an effect on occupational prestige score when comparing to the reference group. Speaking of coefficient, it shows that people with an education level of "below high school", " below high school graduate", "below college graduate", "college graduate and above" has an average occupational prestige score of 33.755, 33.755+ 0.539 =34.294 , 33.755+ 6.433 =40.188 , 33.755+18.61 = 52.365 correspondingly. If R does not have internet access, he/she's occupational prestige score decreases by 3.469.
For the next step, I test the null hypothesis that the effects of all three education levels jointly equal to zero. I can reject the null hypothesis at 5% level of significance, which means that the effects of all three education does not jointly equal to zero. The null hypothesis of being any of the education levels has no occupational prestige score difference can be rejected at the significance level 5%. It means that obtaining a different educational level does have different effects on income. These three education levels do have an effect on the income and therefore improve the model.

Model 2:

```
. xi: regress prestg10 i.educlevel i.intraccess*i.educlevel
i.educlevel       _Ieduclevel_0-3     (naturally coded; _Ieduclevel_0 omitted)
i.intraccess      _Iintracces_0-1     (naturally coded; _Iintracces_0 omitted)
i.int~s*i.edu~l   _IintXedu_#_#       (coded as above)
note: _Ieduclevel_1 omitted because of collinearity.
note: _Ieduclevel_2 omitted because of collinearity.
note: _Ieduclevel_3 omitted because of collinearity.
```

Source	SS	df	MS		
				Number of obs	= 6,290
				F(7, 6282)	= 240.96
Model	331801.777	7	47400.2538	Prob > F	= 0.0000
Residual	1235750.54	6,282	196.712916	R-squared	= 0.2117
				Adj R-squared	= 0.2108
Total	1567552.31	6,289	249.253031	Root MSE	= 14.025

prestg10	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
_Ieduclevel_1	-2.455818	1.749649	-1.40	0.160	-5.885728	.9740912
_Ieduclevel_2	4.24722	1.575045	2.70	0.007	1.159595	7.334846
_Ieduclevel_3	16.33799	1.585236	10.31	0.000	13.23039	19.4456
_Iintracces_1	-6.530791	1.813572	-3.60	0.000	-10.08601	-2.975571
_Ieduclevel_1	0	(omitted)				
_Ieduclevel_2	0	(omitted)				
_Ieduclevel_3	0	(omitted)				
_IintXedu_1_1	4.413948	2.113051	2.09	0.037	.271646	8.556251
_IintXedu_1_2	2.903551	1.8887	1.54	0.124	-.7989474	6.606049
_IintXedu_1_3	3.489085	2.126209	1.64	0.101	-.6790112	7.657181
_cons	35.9878	1.548851	23.24	0.000	32.95153	39.02408

```
. test _IintXedu_1_1 _IintXedu_1_2 _IintXedu_1_3

 ( 1)  _IintXedu_1_1 = 0
 ( 2)  _IintXedu_1_2 = 0
 ( 3)  _IintXedu_1_3 = 0

       F(  3,  6282) =    1.55
            Prob > F =    0.2000

. test _IintXedu_1_1=_IintXedu_1_2=_IintXedu_1_3

 ( 1)  _IintXedu_1_1 - _IintXedu_1_2 = 0
 ( 2)  _IintXedu_1_1 - _IintXedu_1_3 = 0

       F(  2,  6282) =    0.81
            Prob > F =    0.4428
```

Using the interactive model, my reference group remains the same as the additive model. From the P-value shown in regression, I can see that the P-value of the education level variables

changed but the level of “below high school graduate” is still not significant at the significance level of 5%. The P-value for other levels is still <0.05 so I can still conclude that the other education level has an effect on occupational prestige score.
When I compare each interactive variables to the reference group, I can only reject the null hypothesis that being below high school education level without internet access has effects on occupational prestige score at the significance level of 5% because P<0.05; I fail to reject the null hypothesis that being of other education levels without internet has effects on occupational prestige score at the significance level of 5% and 10% since both P-values are over 0.1.
I test the null hypothesis that the effects of being in any of the education levels with no internet access jointly equal to zero. The P-value is 0.2 and I fail to reject the null hypothesis, which means that these variables do not have an effect on the score and do not improve model fit.
I also test the null hypothesis that being in any of the education levels with no internet access has no occupational prestige score difference. The P-value is 0.44 and I fail to reject the null hypothesis, which means that being in any of the education levels with no internet access has no occupational prestige score difference and does not significantly improve the fit of the model.
Speaking of coefficient, the average occupational prestige score for each group is listed below:

“Below high school” and “ internet access at home”:
35.9878
“Below high school” and “no internet access at home”:
35.9878 - 6.53 = 29.4578
“Below high school graduate” and “internet access at home”:
35.9878 - 2.456 = 33.532
“Below high school graduate” and “no internet access at home”:
35.9878 - 2.456 - 6.53 + 4.414 = 31.4158
“Below college graduate” and “internet access at home”
35.9878 + 4.247 = 40.2348
“Below college graduate” and “no internet access at home”
35.9878 + 4.247 - 6.53 + 2.904 = 36.6088
“College graduate and above” and “internet access at home”
35.9878 +16.338 = 52.3258
“College graduate and above” and “no internet access at home”
35.9878 +16.338 - 6.53 + 3.48 = 49.2758

The interactive model illustrates the negative effect of no internet access on occupational prestige score can be alleviated differently at different education levels. For my question, the additive model is a better model than the interactive model, because I fail to reject the null hypothesis that these interactive variables have no effects on the occupational prestige score in t-test.

2) Now let's examine income as a function of education, hours worked per week and sex. Use the same trick for hours worked as in Assignment 4. Now take the natural log of income to better represent its distribution in the U.S. Estimate the model. [First try using the ladder command to see if it concurs with your decision to take the natural log of the variable.] Interpret the coefficients. Now compute the expected natural log of income for men vs. women who have average values on the other variables. [Try graphing in Stata or some other program if you're feeling ambitious.]

```
. ladder incomenew if good ==1
```

Transformation	Formula	chi2(2)	Prob > chi2
Cubic	income~w^3	**409.17**	**0.000**
Square	income~w^2	**292.27**	**0.000**
Identity	income~w	**94.32**	**0.000**
Square root	sqrt(income~w)	**29.21**	**0.000**
Log	log(income~w)	**219.58**	**0.000**
1/(Square root)	1/sqrt(income~w)	**1043.94**	**0.000**
Inverse	1/income~w	.	.
1/Square	1/(income~w^2)	.	.
1/Cubic	1/(income~w^3)	.	.

```
. regress incomelog educ hrs_wrked i.sex if good==1
```

Source	SS	df	MS	Number of obs	=	1,147
				F(3, 1143)	=	78.36
Model	161.345773	3	53.7819243	Prob > F	=	0.0000
Residual	784.507105	1,143	.686357922	R-squared	=	0.1706
				Adj R-squared	=	0.1684
Total	945.852878	1,146	.825351551	Root MSE	=	.82847

incomelog	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ	.1137286	.0083822	13.57	0.000	.0972823	.1301748
hrs_wrked	.0113069	.0017418	6.49	0.000	.0078895	.0147244
sex						
female	-.0858804	.0506684	-1.69	0.090	-.1852939	.0135332
_cons	8.78305	.1431586	61.35	0.000	8.502167	9.063933

This is a log-level model. My reference group is male. Speaking of coefficients, a one-unit increase in education increases the income by 11.37%, a one-unit increase in hours worked increases the income by 1.13%, and being a female causes income to decrease by 8%. Only education and hours worked are statistically significant at the significance level of 5%.

Expected natural log of income of an average female is:

```
. mean educ if sex==2 & good==1

Mean estimation                         Number of obs = 543
```

	Mean	Std. err.	[95% conf.	interval]
educ	13.96133	.1209895	13.72366	14.19899

```
. mean hrs_wrked if sex==2 & good==1

Mean estimation                         Number of obs = 543
```

	Mean	Std. err.	[95% conf.	interval]
hrs_wrked	38.20994	.5507239	37.12813	39.29176

incomelog = 8.783 + 0.114*educ + 0.011*hrs_wrked - 0.086
incomelog = 8.783 + 0.114*13.961 + 0.011*38.2 - 0.086 = 10.708754

Expected natural log of income of an average male is:

```
. mean educ if sex==1 & good==1

Mean estimation                         Number of obs = 604
```

	Mean	Std. err.	[95% conf.	interval]
educ	13.69205	.1224469	13.45158	13.93253

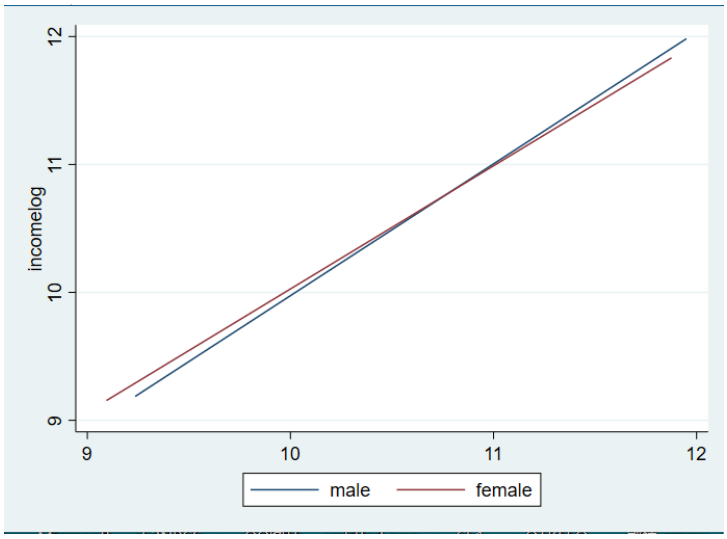
```
. mean hrs_wrked if sex==1 & good==1

Mean estimation                         Number of obs = 604
```

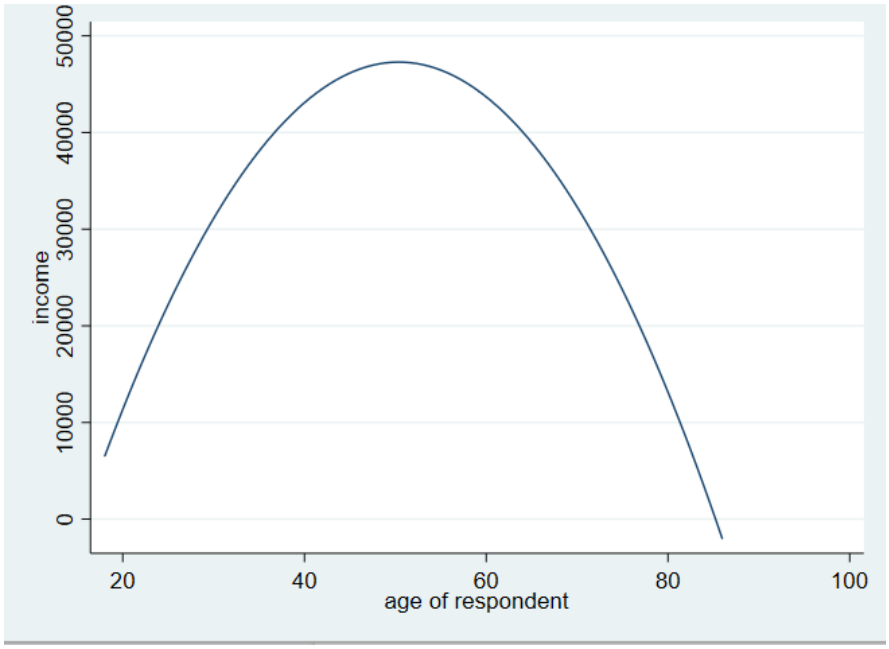
	Mean	Std. err.	[95% conf.	interval]
hrs_wrked	45.50662	.6132718	44.30221	46.71103

incomelog = 8.783 + 0.114*educ + 0.011*hrs_wrked
incomelog = 8.783 + 0.114*13.692 + 0.011*45.51 = 10.844

The graph below illustrates the formulas used above:



3) Using gss_2004, examine the curvilinear relationship between age and income. Begin by recoding "rincom98" in the way we've become familiar with in this course (midpoint of each interval). Take advantage of previous cross-sectional research that demonstrates the validity of a quadratic term for age (that is, age2) and estimate a model including age and age2 as predictors of income. Try to interpret the relationship. Predicted values or a graph might be useful here. Be ambitious and try graphing! Speculate about a potential reason for the curvilinear relationship.



By graphing the quadratic prediction of incomenew using age and agesqr, I see that there's a downturn around age 50 in the positive correlation between income and age. I then proceed to perform a regression analysis with a quadratic model.

```
. regress incomenew age agesqr if good==1
```

Source	SS	df	MS		
Model	1.7757e+11	2	8.8784e+10	Number of obs	= 1,685
Residual	1.3618e+12	1,682	809647592	F(2, 1682)	= 109.66
				Prob > F	= 0.0000
Total	1.5394e+12	1,684	914129723	R-squared	= 0.1153
				Adj R-squared	= 0.1143
				Root MSE	= 28454

incomenew	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	3922.005	301.9133	12.99	0.000	3329.84	4514.17
agesqr	-38.9185	3.365293	-11.56	0.000	-45.5191	-32.3179
_cons	-51524.97	6393.085	-8.06	0.000	-64064.21	-38985.73

Since age, the independent variable, is squared, it is a quadratic model.
All p-values on the table are less than 0.05, which means that we are able to reject the null hypothesis that age and agesqr do not have effects on incomenew at the significance level of 5%.
R-square indicates that 11.53% of the variables are explained by age and agesqr.

Using the coefficient values in the regression table, the equation would be:
income = -51524 + 3922*age - 38.9185*agesqr
This equation is helpful in predicting the average expected income of people at a certain age.
Below are the samples:

Age 20:
income = -51524 + 3922*20 - 38.9185*20^2 = 11348.6
Age 40:
income = -51524 + 3922*40 - 38.9185*40^2 = 43086.4
Age 70:
income = -51524 + 3922*70 - 38.9185*70^2 = 32315.35

From the predicted values, it is clear that there's a curvilinear relationship between income and age. To further interpret that, it means that as age increases, its relationship with income changes. They are initially positively correlated but change to negatively correlated at approximately around 50 years old. A possible explanation is that when people retire around 50, their income decreases.