Data Analysis with Python

Log Transformations + Interactions + More on Multiple Regression

(Class #4)

Gregory M. Eirich
QMSS

Agenda

1. Log transformations
2. Interactions

1. Log transformations

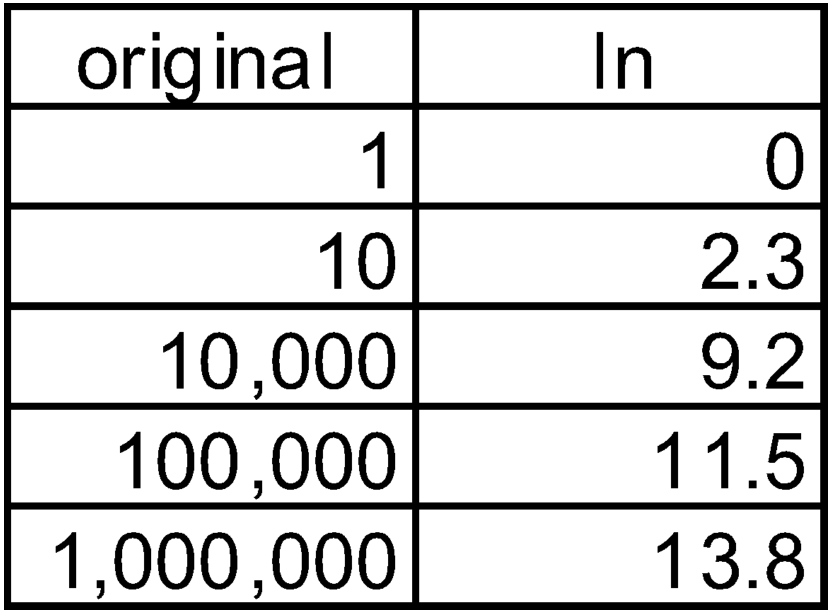
Why Log Transform Variables?

Log Transformations

In order to make some variables “more normal,” or more linear, or to increase interpretability, we often log them.

Logging

The natural logarithm of a number is the exponent to which we have to raise the *base(~2.72)* to obtain *that number*



original	ln
1	0
10	2.3
10,000	9.2
100,000	11.5
1,000,000	13.8

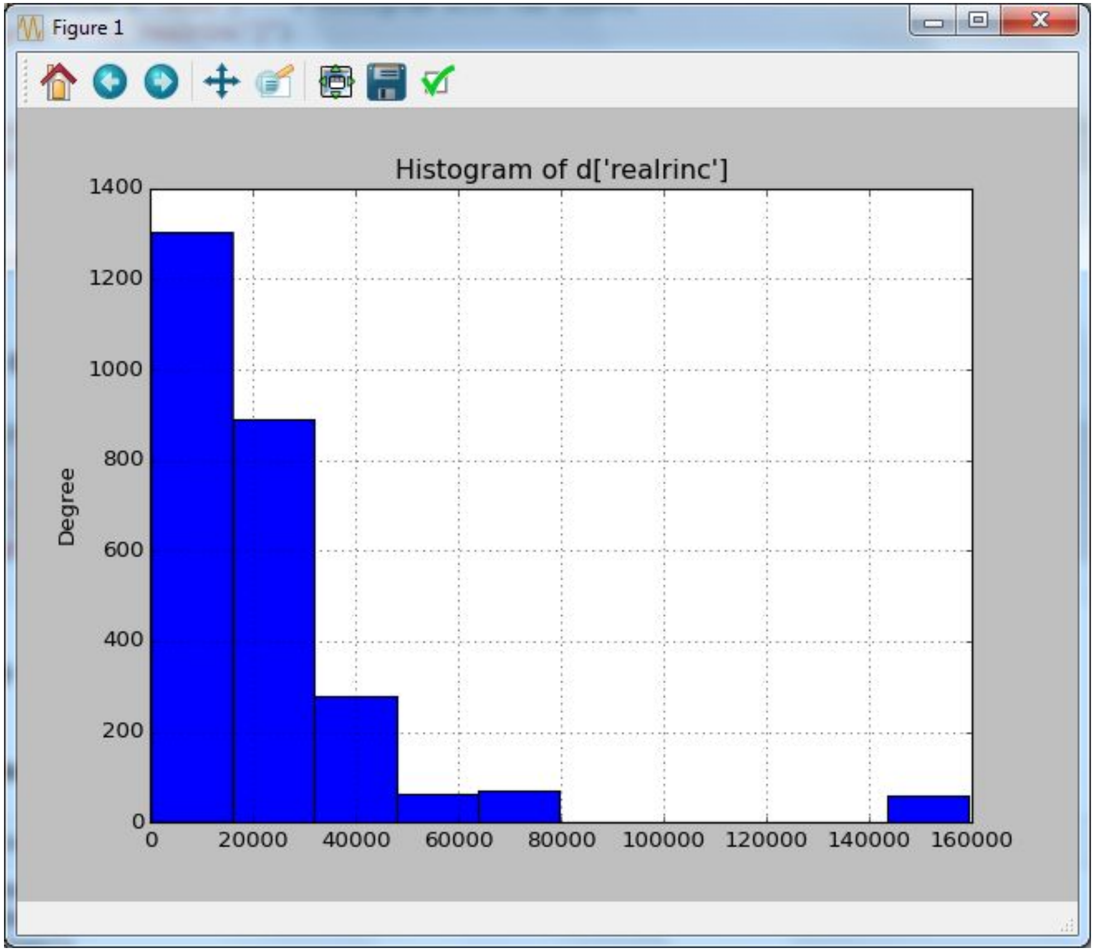
N.B., You cannot take the log of 0 … this can be a problem

Preliminaries...

```
from __future__ import division
import pandas as pd
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
import os
import matplotlib.pyplot as plt


os.chdir('C:/Users/gme2101/Desktop/Data Analysis Data') # change working directory
d = pd.read_csv("GSS_Cum.csv", usecols=["sex", "educ", "year", "realrinc", "hrs1", "wordsum",
"wrkstat", "race", "trust", "region", "fund", "evolved", "realinc", "sibs", "madeg", "fund",
"marital", "attend", "age", "family16"])
d.head()

sub = d[d["year"] == 2006]
```

Distribution of *realrinc*

```
sub["realrinc"].plot(kind
= 'hist')    # histogram
with raw counts
plt.title("Histogram of
d['realrinc']")
plt.show()
```

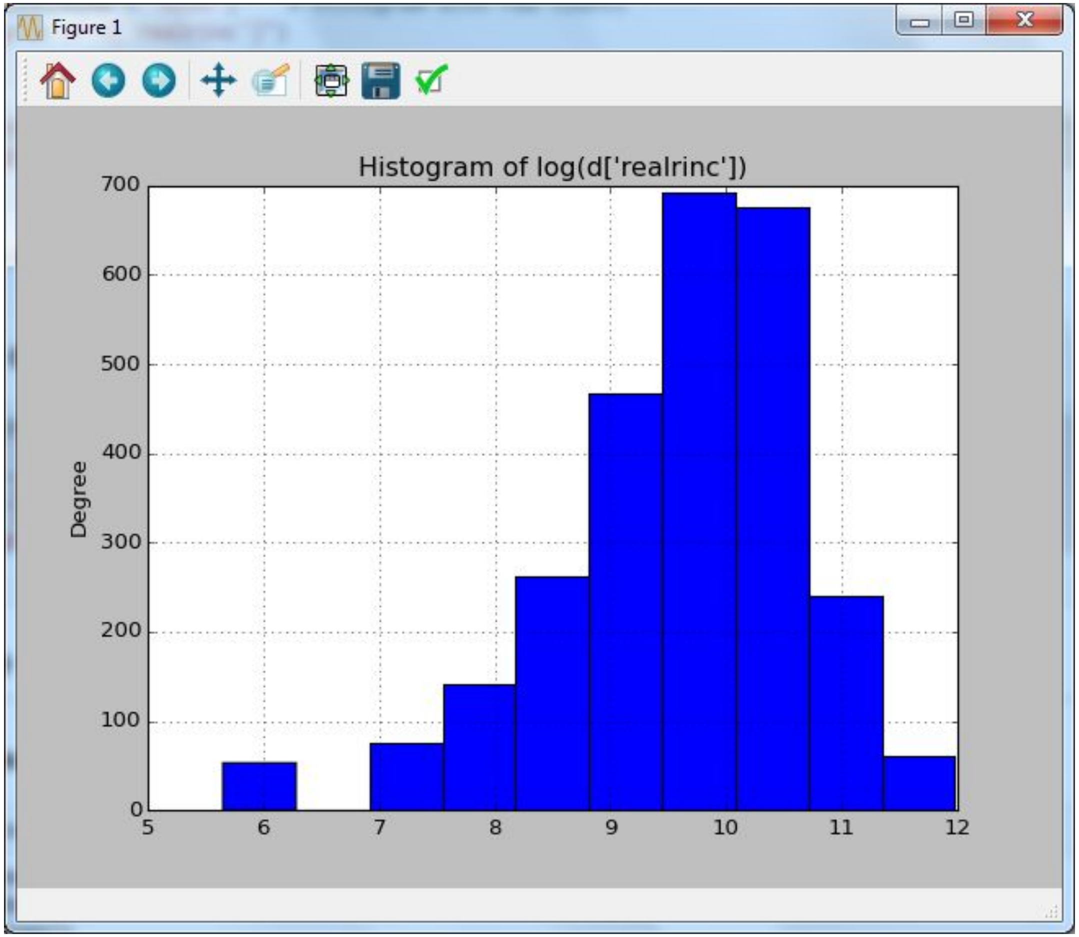


(c) Eirich 2013

*

Distribution of *ln.realrinc*

```
pd.options.mode.chained_assignment = None

sub["ln_realrinc"] = np.log(sub["realrinc"])
sub["ln_realrinc"].plot(kind = 'hist')
plt.title("Histogram of log(d['realrinc'])")
plt.show()
```



*

Looking at the shape of our variables

```
## RAW income ##

sub["realrinc"].skew()

3.4025770427387112

sub["realrinc"].kurtosis()

14.553164088336334
```

Raw income has skew=3.4 and kurtosis=14.51, while a normal, symmetric distribution will have skew=0, kurtosis=3;

```
## LOGGED income ##

sub["ln_realrinc"].skew()

-1.0295369876382472

sub["ln_realrinc"].kurtosis()

2.0276359476609085
```

Log(income) has skew of -1.03 and kurtosis of 2.02.

Which is closer to our ideal normal distribution?

Log Transformations

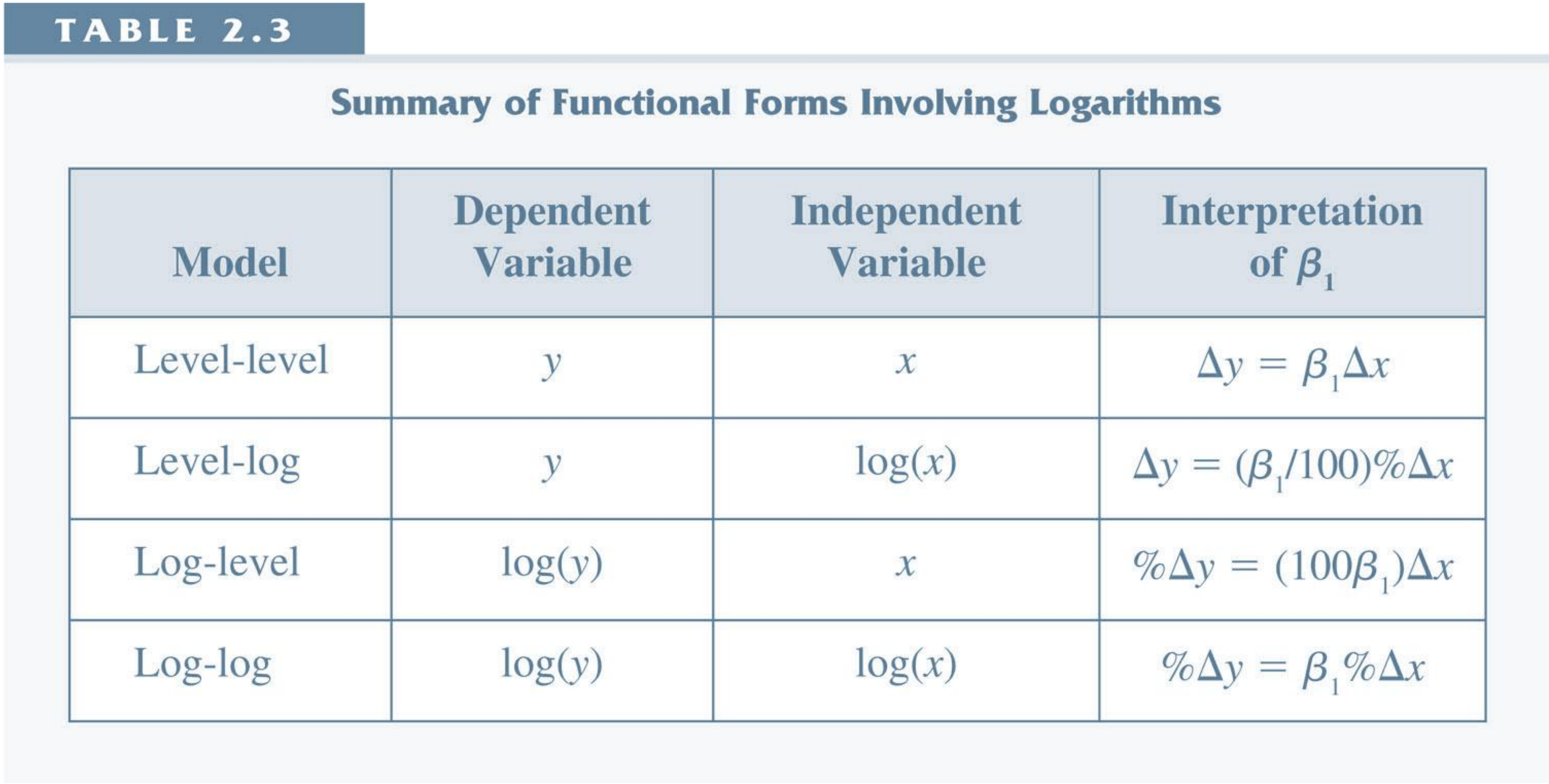


TABLE 2.3

Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

A log-log model

```
pd.options.mode.chained_assignment = None

sub["ln_hrs1"] = np.log(sub["hrs1"])
sub["ln_realrinc"] = np.log(sub["realrinc"])
lm1 = smf.ols(formula = "ln_realrinc ~ ln_hrs1 + C(sex)", data = sub).fit()
print (lm1.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:            ln_realrinc   R-squared:                       0.181
Model:                            OLS   Adj. R-squared:                  0.180
Method:                 Least Squares   F-statistic:                     250.6
Date:                Wed, 07 Jun 2017   Prob (F-statistic):           4.60e-99
Time:                        16:15:49   Log-Likelihood:                -3070.9
No. Observations:                2275   AIC:                             6148.
Df Residuals:                    2272   BIC:                             6165.
Df Model:                           2
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------
Intercept       6.6776      0.176     38.049      0.000         6.333     7.022
C(sex)[T.2]    -0.3383      0.040     -8.479      0.000        -0.417    -0.260
ln_hrs1         0.8645      0.046     18.732      0.000         0.774     0.955
==============================================================================
Omnibus:                      434.513   Durbin-Watson:                   1.792
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1354.151
Skew:                          -0.962   Prob(JB):                    8.91e-295
Kurtosis:                       6.253   Cond. No.                         35.9
==============================================================================

Warnings:
```

A log-log model

```
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      6.6776      0.176     38.049      0.000         6.333     7.022
C(sex)[T.2]   -0.3383      0.040     -8.479      0.000        -0.417    -0.260
ln_hrs1        0.8645      0.046     18.732      0.000         0.774     0.955
=============================================================================
```

Controlling for sex, a 1% increase in work hours leads (on average) to a 0.86% increase in salary

A level-log model

```
pd.options.mode.chained_assignment = None

sub["ln_wordsum"] = np.log(sub["wordsum"])
sub["working"] = sub["wrkstat"].apply(lambda e: 1 if e < 3 else 0)
sub["ln_wordsum"] = sub["ln_wordsum"].map(lambda x: np.nan if x == -float('Inf') else x)
lm2 = smf.ols(formula = "tvhours ~ ln_wordsum + working", data = sub).fit()
print (lm2.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                tvhours   R-squared:                       0.075
Model:                            OLS   Adj. R-squared:                  0.073
Method:                 Least Squares   F-statistic:                     37.34
Date:                Mon, 20 May 2019   Prob (F-statistic):           2.57e-16
Time:                        15:29:30   Log-Likelihood:                -2080.8
No. Observations:                 921   AIC:                             4168.
Df Residuals:                     918   BIC:                             4182.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      5.4222      0.360     15.046      0.000         4.715     6.130
ln_wordsum    -1.0379      0.194     -5.360      0.000        -1.418    -0.658
working       -1.0174      0.155     -6.571      0.000        -1.321    -0.714
==============================================================================
Omnibus:                      556.467   Durbin-Watson:                   2.053
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7021.934
Skew:                           2.543   Prob(JB):                         0.00
Kurtosis:                      15.535   Cond. No.                         11.4
==============================================================================
```

A level-log model

```
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      5.4222      0.360     15.046      0.000         4.715     6.130
ln_wordsum    -1.0379      0.194     -5.360      0.000        -1.418    -0.658
working       -1.0174      0.155     -6.571      0.000        -1.321    -0.714
==============================================================================
```

Controlling for working status, a 1% increase in vocabulary score leads (on average) to a -0.0104 hour decrease in TV hours

*

A level-log model

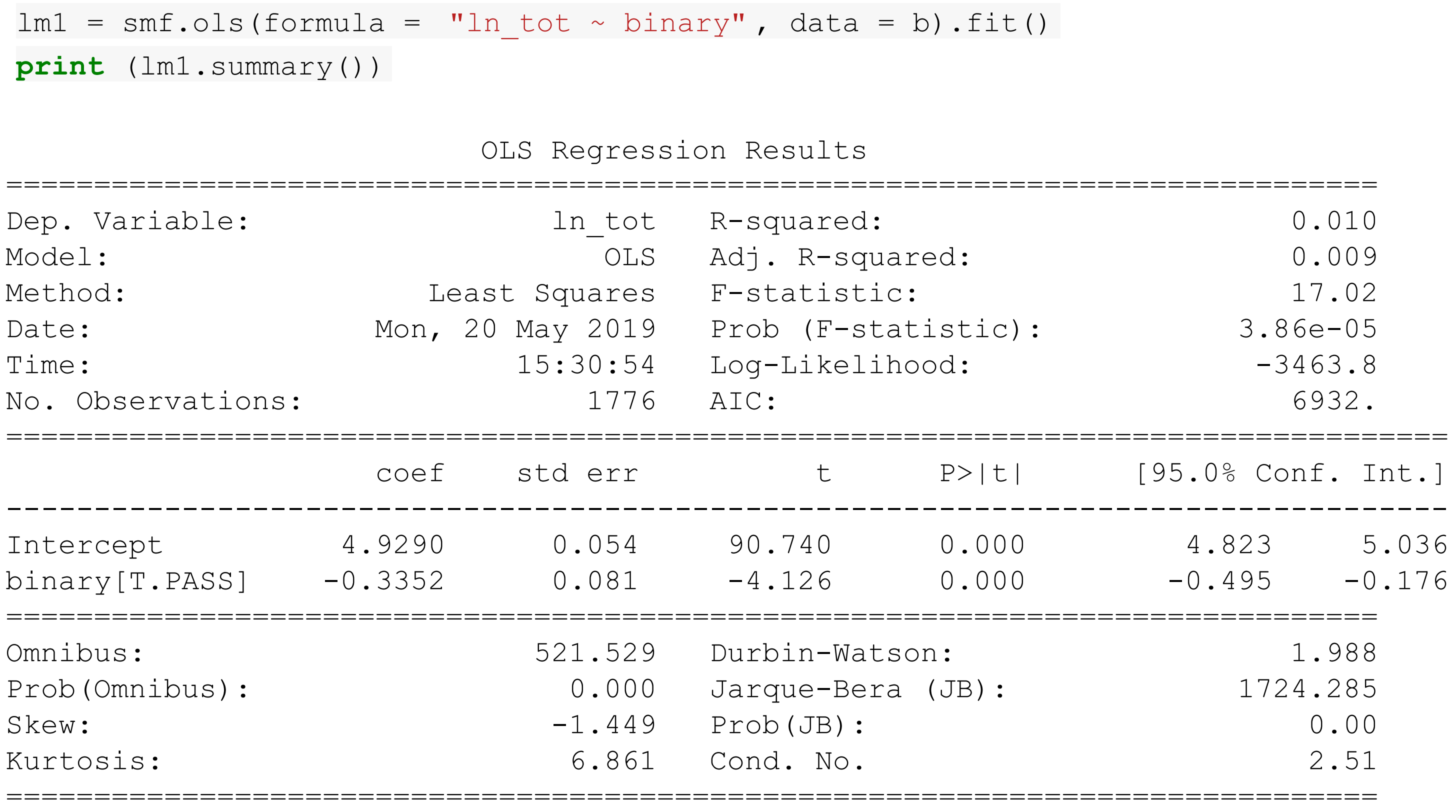
```
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      5.4222      0.360     15.046      0.000         4.715     6.130
ln_wordsum    -1.0379      0.194     -5.360      0.000        -1.418    -0.658
working       -1.0174      0.155     -6.571      0.000        -1.321    -0.714
==============================================================================
```

Or: Controlling for working status, a 100% increase in vocabulary score leads (on average) to a 1.04 hour decrease in TV hours*

Another log-log model

```
b["tg13"] = b["domgross_2013$"] + b["intgross_2013$"]
b["tot_gross_13_mil"] = b["tg13"] / (1000000)
b["budget_13_mil"] = b["budget_2013$"] / (1000000)

b["ln_bud"] = np.log(b["budget_13_mil"])
b["ln_tot"] = np.log(b["tot_gross_13_mil"])
```

Another log-log model

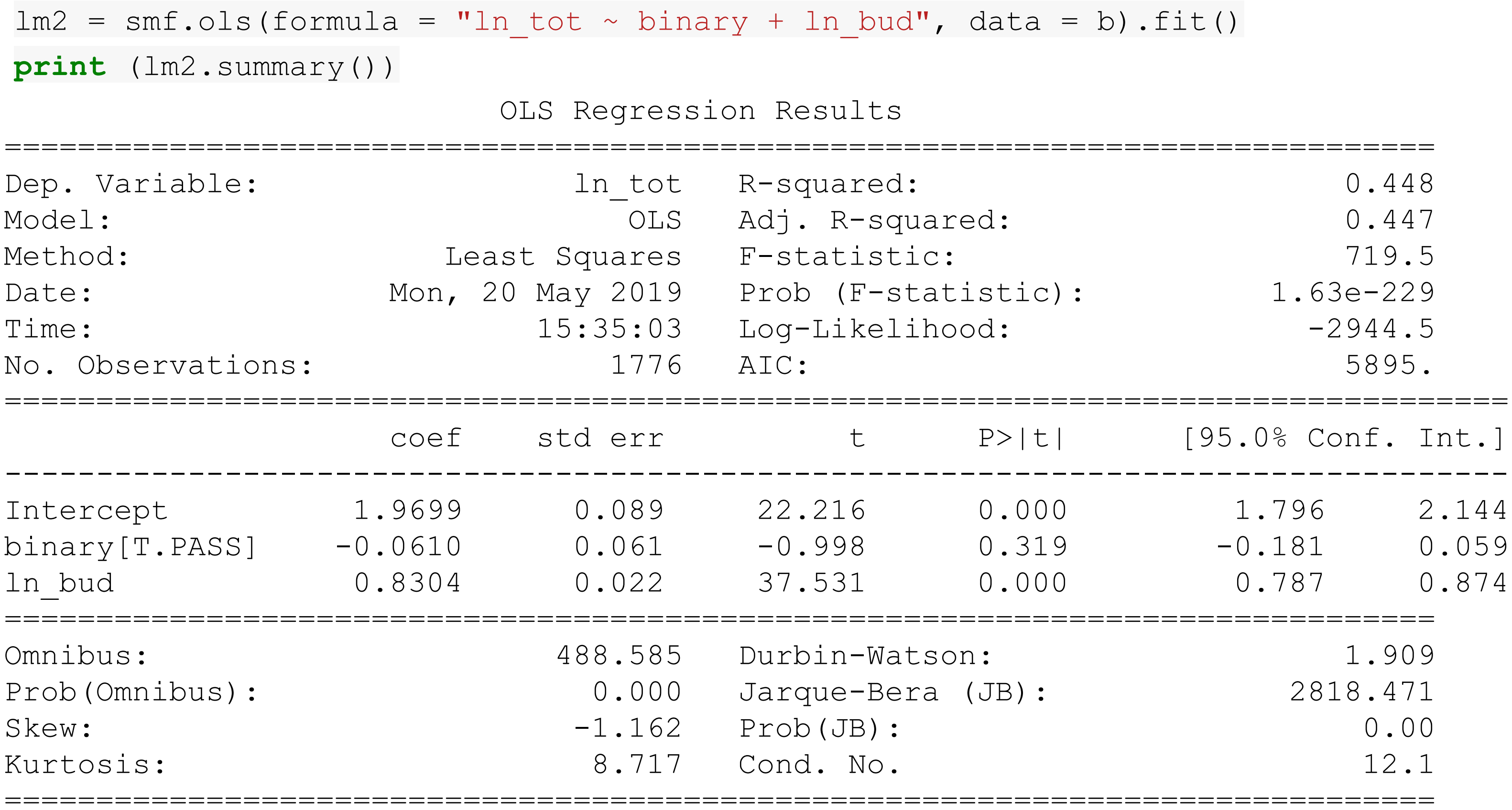


```
lm1 = smf.ols(formula = "ln_tot ~ binary", data = b).fit()
print (lm1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 ln_tot   R-squared:                       0.010
Model:                            OLS   Adj. R-squared:                  0.009
Method:                 Least Squares   F-statistic:                     17.02
Date:                Mon, 20 May 2019   Prob (F-statistic):           3.86e-05
Time:                        15:30:54   Log-Likelihood:                -3463.8
No. Observations:                1776   AIC:                             6932.
==================================================================================
                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Intercept          4.9290      0.054     90.740      0.000         4.823     5.036
binary[T.PASS]    -0.3352      0.081     -4.126      0.000        -0.495    -0.176
==============================================================================
Omnibus:                      521.529   Durbin-Watson:                   1.988
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1724.285
Skew:                          -1.449   Prob(JB):                         0.00
Kurtosis:                       6.861   Cond. No.                         2.51
==============================================================================
```

Model 1: Passing the Bechler test reduces the predicted total revenues of a movie by 33.5%

Another log-log model



```
lm2 = smf.ols(formula = "ln_tot ~ binary + ln_bud", data = b).fit()
print (lm2.summary())
                           OLS Regression Results
==============================================================================
Dep. Variable:                 ln_tot   R-squared:                       0.448
Model:                            OLS   Adj. R-squared:                  0.447
Method:                 Least Squares   F-statistic:                     719.5
Date:                Mon, 20 May 2019   Prob (F-statistic):          1.63e-229
Time:                        15:35:03   Log-Likelihood:                -2944.5
No. Observations:                1776   AIC:                             5895.
==================================================================================
                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Intercept          1.9699      0.089     22.216      0.000         1.796     2.144
binary[T.PASS]    -0.0610      0.061     -0.998      0.319        -0.181     0.059
ln_bud             0.8304      0.022     37.531      0.000         0.787     0.874
==============================================================================
Omnibus:                      488.585   Durbin-Watson:                   1.909
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2818.471
Skew:                          -1.162   Prob(JB):                         0.00
Kurtosis:                       8.717   Cond. No.                         12.1
==============================================================================
```

Model 2: Controlling for the Bechler test, a 1% increase in the budget of a movie increases its predicted revenues by 0.83%

Another log-log model

```
==================================================================================
                       coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Intercept            1.9699      0.089     22.216      0.000         1.796     2.144
binary[T.PASS]      -0.0610      0.061     -0.998      0.319        -0.181     0.059
ln_bud               0.8304      0.022     37.531      0.000         0.787     0.874
==============================================================================
```

Model 2: Controlling for budget, a movie that passes the Bechdel test is predicted to reduce revenues by 6.1% on average (n.s.)

Remember the original model

```
lm3 = smf.ols(formula = "tot_gross_13_mil ~ binary + budget_13_mil", data = b).fit()
print (lm3.summary()) --      OLS Regression Results
==============================================================================
Dep. Variable:       tot_gross_13_mil   R-squared:                       0.316
Model:                            OLS   Adj. R-squared:                  0.315
Method:                 Least Squares   F-statistic:                     408.7
Date:                Mon, 20 May 2019   Prob (F-statistic):          1.10e-146
Time:                        15:35:07   Log-Likelihood:                -12839.
No. Observations:                1776   AIC:                         2.568e+04
==================================================================================
                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Intercept         71.4731     14.099      5.069      0.000        43.820    99.126
binary[T.PASS]   -14.9222     16.123     -0.926      0.355       -46.543    16.699
budget_13_mil      4.0963      0.146     28.108      0.000         3.810     4.382
==============================================================================
Omnibus:                     1696.847   Durbin-Watson:                   1.942
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           104439.408
Skew:                           4.389   Prob(JB):                         0.00
Kurtosis:                      39.528   Cond. No.                         190.
```

Another log-log model

```
Regression Results
=====================================================
                   Ln(Total Gross)   Total Gross, Mil
                   Model 1  Model 2      Model 3
-----------------------------------------------------
Pass Bechler      -0.335***  -0.061      -14.922
                   (0.081)  (0.061)      (16.123)
Ln(Total Budget)            0.830***
                            (0.022)
Total Budget, Mil                        4.096***
                                         (0.146)
Constant          4.929***  1.970***    71.473***
                   (0.054)  (0.089)      (14.099)
-----------------------------------------------------
Observations        1,776    1,776        1,776
Adjusted R2         0.009    0.447        0.315
=====================================================
Note:                     *p<0.1; **p<0.05; ***p<0.01
```

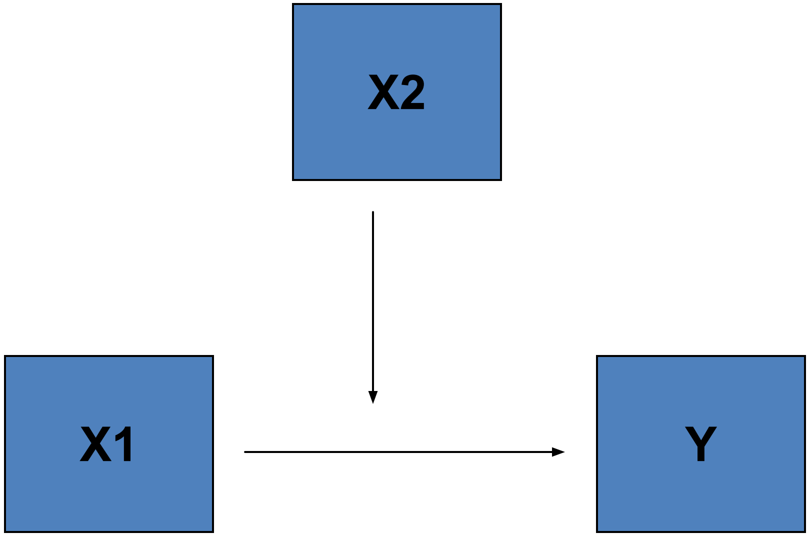
How’d I do that (in R)?

```
lm1 = lm(d$ln.tot ~ binary , d)
lm2 = lm(d$ln.tot ~ binary + ln.bud, d)
lm3 = lm(d$tot.gross.13.mil ~ binary + d$budget.13.mil, d)

library(stargazer)
stargazer(lm1, lm2, type = "text")

stargazer(lm1, lm2, lm3,
          title="Regression Results",
          align=TRUE,
          dep.var.labels=c("Ln(Total Gross)",  "Total Gross, Mil"),
          covariate.labels=c("Pass Bechler","Ln(Total Budget)", "Total 
Budget, Mil"),
          no.space=TRUE,
          omit.stat=c("LL","ser","f", "rsq"),
          column.labels=c("Model 1", "Model 2", "Model 3"),
          dep.var.caption="",
          model.numbers=FALSE,
          type = "text")
```

2. Interactions

Interactions

- Moderation: There is an “interaction” between our X1 and X2
- The association of X1 and Y varies according to levels of X2



Why would be use interactions?

If we think that 2 of our independent variables may have a relationship that affects the magnitude of our dependent variable

- That the effect of 1 variable may depend on the magnitude of another variable …

Moderation: Our simple model

$$Y = a + B_1X_1 + B_2X_2 + u$$

$$\text{R's Income} = a + B_1(\text{Gender}) + B_2(\text{Educ})$$

Why might gender and education interact to predict salary?

At higher levels of education, are the differences amplified or minimized between males and females on income?

These are the recodes …
Income in $10,000 units

```
sub_new = sub[["realrinc", "educ", "sex"]]
sub_new["female"] = sub_new["sex"] == 2
sub_new.dropna(subset = ["realrinc", "educ"], inplace = True)
sub_new["realrinc10k"] = (sub_new.realrinc) /10000
sub_new["realrinc10k"].describe()
```

```
count    2663.00
mean        2.36
std         2.60
min         0.03
25%         0.92
50%         1.85
75%         3.13
max        15.93
Name: realrinc10k, dtype: float64
```

A simple multiple regression

```
pd.options.display.float_format = '{0:1.2f}'.format
lm_income = smf.ols(formula = "realrinc10k ~ educ + female", data = sub_new).fit()
print (lm_income.summary())
sub_new["fitted"] = lm_income.predict()
```

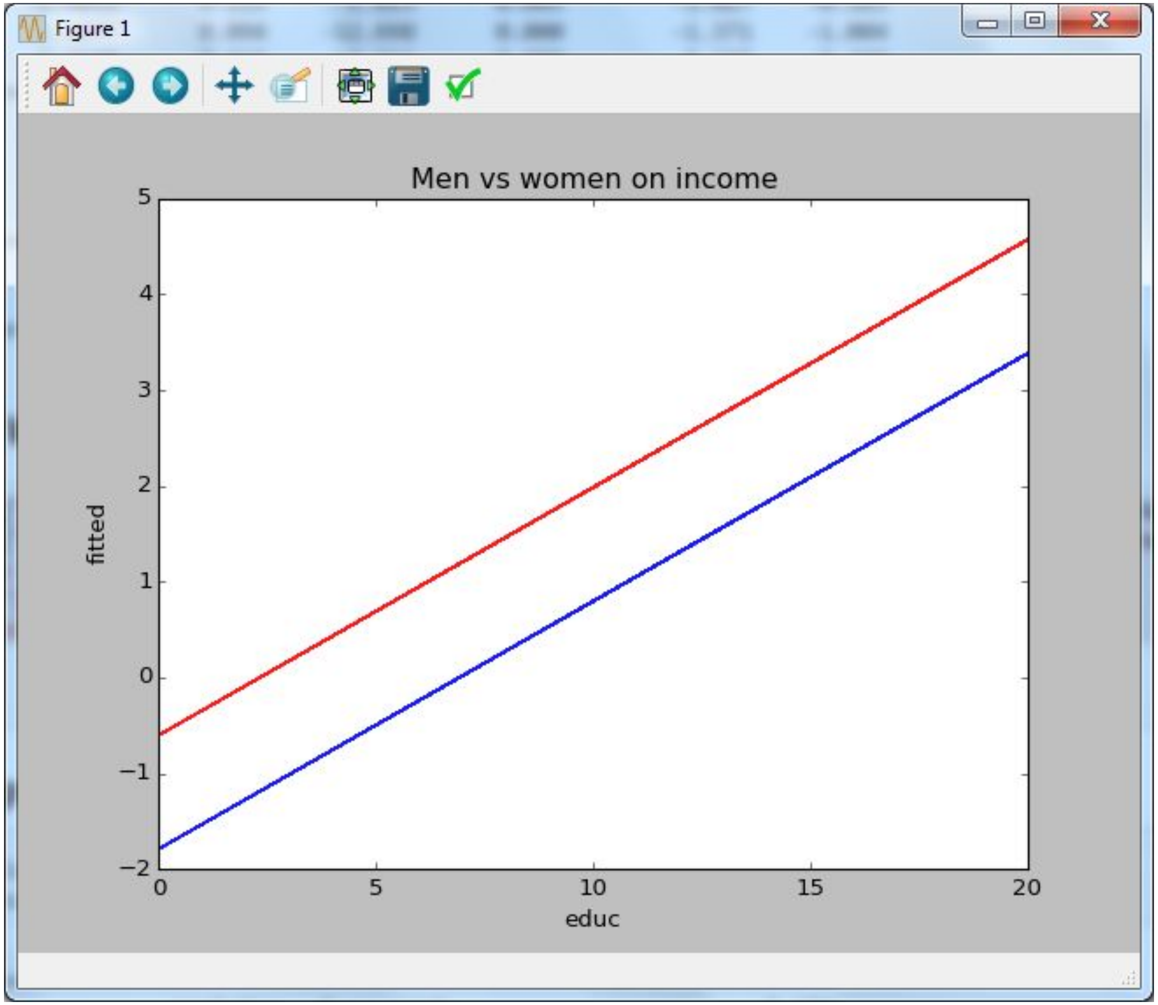
```
                           OLS Regression Results
==============================================================================
Dep. Variable:            realrinc10k   R-squared:                       0.142
Model:                            OLS   Adj. R-squared:                  0.141
Method:                 Least Squares   F-statistic:                     220.1
Date:                Tue, 21 May 2019   Prob (F-statistic):           3.64e-89
Time:                        09:28:39   Log-Likelihood:                -6121.0
No. Observations:                2663   AIC:                         1.225e+04
Df Residuals:                    2660   BIC:                         1.227e+04
Df Model:                           2
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Intercept         -0.6051      0.215     -2.813      0.005        -1.027    -0.183
female[T.True]    -1.1877      0.094    -12.698      0.000        -1.371    -1.004
educ               0.2588      0.015     17.211      0.000         0.229     0.288
==============================================================================
Omnibus:                     1902.623   Durbin-Watson:                   1.890
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            29109.520
Skew:                           3.292   Prob(JB):                         0.00
Kurtosis:                      17.799   Cond. No.                         65.4
==============================================================================
```

A simple multiple regression

For each additional year of education, net of sex, someone earns $2,588 more (statistically significant) per year; a female – net of education – earns $11,877 less per year (statistically significant)

```
===================================================================================
                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------
Intercept         -0.6051      0.215     -2.813      0.005        -1.027    -0.183
female[T.True]    -1.1877      0.094    -12.698      0.000        -1.371    -1.004
educ               0.2588      0.015     17.211      0.000         0.229     0.288
===============================================================================
```

Female vs. male on income



How did I do that graph?

```
plt.plot(sub_new["educ"], lm_income.params[0] + lm_income.params[1] * 1 +
lm_income.params[2] * sub_new["educ"], 'b', label = 'female', alpha = 0.9)
plt.plot(sub_new["educ"], lm_income.params[0] + lm_income.params[1] * 0 +
lm_income.params[2] * sub_new["educ"], 'r', label = 'male', alpha = 0.9)
plt.title("Men vs women on income")
plt.xlabel("educ")
plt.ylabel("fitted")
plt.show()
```

A simple regression for males

For males, each additional year of education, they earn $2,990 per year (statistically significant)

(A male with no education (X=0) has -$11,522)

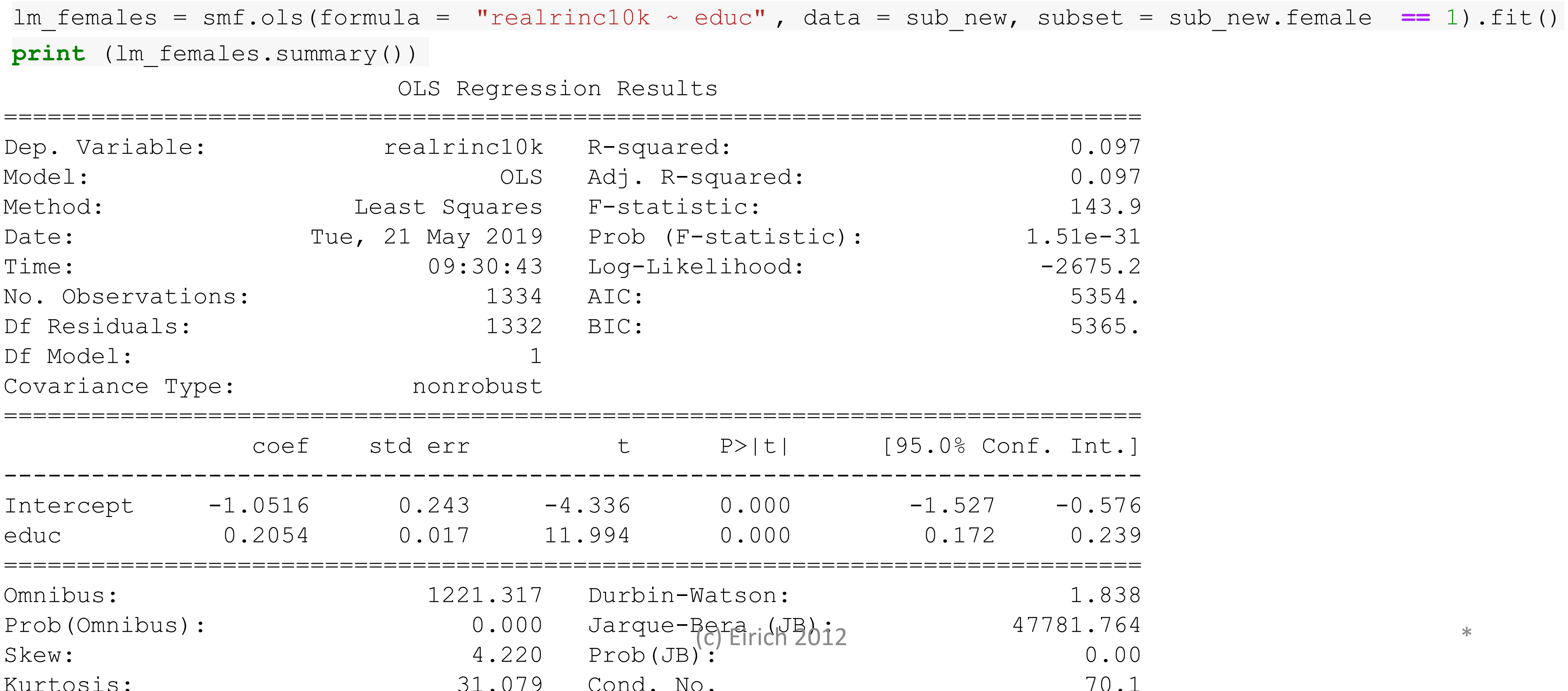
```
lm_males = smf.ols(formula = "realrinc10k ~ educ", data = sub_new, subset = sub_new.female == 0).fit()
print (lm_males.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            realrinc10k   R-squared:                       0.106
Model:                            OLS   Adj. R-squared:                  0.105
Method:                 Least Squares   F-statistic:                     156.9
Date:                Tue, 21 May 2019   Prob (F-statistic):           4.30e-34
Time:                        09:30:22   Log-Likelihood:                -3296.2
No. Observations:                1329   AIC:                             6596.
Df Residuals:                    1327   BIC:                             6607.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     -1.1522      0.335     -3.444      0.001        -1.808    -0.496
educ           0.2990      0.024     12.524      0.000         0.252     0.346
==============================================================================
Omnibus:                      796.985   Durbin-Watson:                   1.918
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6573.585
Skew:                           2.760   Prob(JB):                         0.00
Kurtosis:                      12.394   Cond. No.                         59.4
```

A simple regression for females

For females, each additional year of education, they earn $2,053 per year (statistically significant)

(A female with no education (X=0) earns -$10,516 per year)



```
lm_females = smf.ols(formula =  "realrinc10k ~ educ" , data = sub_new, subset = sub_new.female  == 1).fit()
print (lm_females.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            realrinc10k   R-squared:                       0.097
Model:                            OLS   Adj. R-squared:                  0.097
Method:                 Least Squares   F-statistic:                     143.9
Date:                Tue, 21 May 2019   Prob (F-statistic):           1.51e-31
Time:                        09:30:43   Log-Likelihood:                -2675.2
No. Observations:                1334   AIC:                             5354.
Df Residuals:                    1332   BIC:                             5365.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     -1.0516      0.243     -4.336      0.000        -1.527    -0.576
educ           0.2054      0.017     11.994      0.000         0.172     0.239
==============================================================================
Omnibus:                     1221.317   Durbin-Watson:                   1.838
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            47781.764
Skew:                           4.220   Prob(JB):                         0.00
Kurtosis:                      31.079   Cond. No.                         70.1
```

The interaction model I

1. When female=0 (i.e., for males), at the intercept, they earn -$11,522 on average with 0 years of education.

```
# Note: the * in the formula means that we want the interaction term in addition to each term separately
# Note: use ':' instead if you want to include the interaction term only
lm_income2 = smf.ols(formula =  "realrinc10k ~ educ * female" , data = sub_new).fit()
print (lm_income2.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:            realrinc10k   R-squared:                       0.145
Model:                            OLS   Adj. R-squared:                  0.144
Method:                 Least Squares   F-statistic:                     150.4
Date:                Tue, 21 May 2019   Prob (F-statistic):           5.30e-90
Time:                        09:31:21   Log-Likelihood:                -6116.2
No. Observations:                2663   AIC:                         1.224e+04
Df Residuals:                    2659   BIC:                         1.226e+04
Df Model:                           3
Covariance Type:            nonrobust
=========================================================================================
                            coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------------
Intercept              -1.1522      0.278     -4.138      0.000        -1.698    -0.606
female[T.True]          0.1006      0.428      0.235      0.814        -0.738     0.939
educ                    0.2990      0.020     15.046      0.000         0.260     0.338
educ:female[T.True]    -0.0936      0.030     -3.087      0.002        -0.153    -0.034
==============================================================================
Omnibus:                     1893.429   Durbin-Watson:                   1.889
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            28848.466
Skew:                           3.270   Prob(JB):                         0.00
```

The interaction model II

2. When female=0 (i.e., for males), they earn $2,990 on average for each additional year of education.

```
======================================================================================
                          coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------------
Intercept              -1.1522      0.278     -4.138      0.000        -1.698    -0.606
female[T.True]          0.1006      0.428      0.235      0.814        -0.738     0.939
educ                    0.2990      0.020     15.046      0.000         0.260     0.338
educ:female[T.True]    -0.0936      0.030     -3.087      0.002        -0.153    -0.034
==============================================================================
```

The interaction model III

3. When educ=0, for females (female=1), they earn $1006 on average more than males.

```
=======================================================================================
                          coef     std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------------
Intercept              -1.1522       0.278     -4.138      0.000        -1.698    -0.606
female[T.True]          0.1006       0.428      0.235      0.814        -0.738     0.939
educ                    0.2990       0.020     15.046      0.000         0.260     0.338
educ:female[T.True]    -0.0936       0.030     -3.087      0.002        -0.153    -0.034
===============================================================================
```

The interaction model IV

4a. For females, they get $936 less than males for each year more of education, so males get $2,990, but females get $2,990 - $936 = $2054, on average

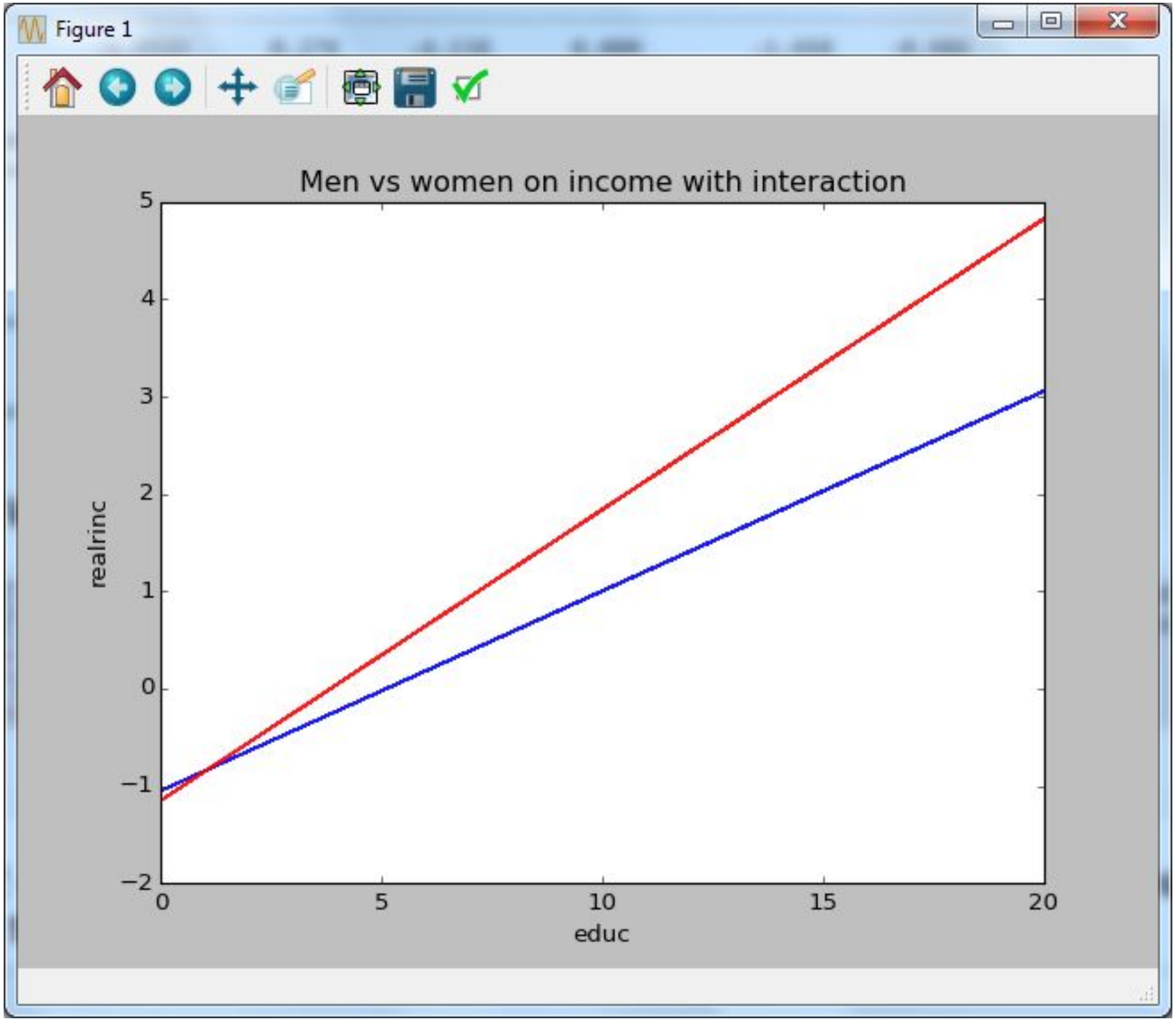
```
=======================================================================================
                          coef    std err          t      P>|t|       [95.0% Conf. Int.]
----------------------------------------------------------------------------------------
Intercept              -1.1522      0.278     -4.138      0.000        -1.698     -0.606
female[T.True]          0.1006      0.428      0.235      0.814        -0.738      0.939
educ                    0.2990      0.020     15.046      0.000         0.260      0.338
educ:female[T.True]    -0.0936      0.030     -3.087      0.002        -0.153     -0.034
===============================================================================
```

The interaction model IV

4b. Or: For females on average, the *difference* in rates of return to education is $936 per year less than for males

```
=======================================================================================
                          coef    std err          t      P>|t|       [95.0% Conf. Int.]
----------------------------------------------------------------------------------------
Intercept              -1.1522      0.278     -4.138      0.000         -1.698    -0.606
female[T.True]          0.1006      0.428      0.235      0.814         -0.738     0.939
educ                    0.2990      0.020     15.046      0.000          0.260     0.338
educ:female[T.True]    -0.0936      0.030     -3.087      0.002         -0.153    -0.034
===============================================================================
```

Male vs. female … with interaction



*

How did I do this graph?

```
plt.plot(sub_new["educ"], lm_income2.params[0] + lm_income2.params[1] * 1 +
lm_income2.params[2] * sub_new["educ"] + lm_income2.params[3] * 1 *
sub_new["educ"], 'b', label = 'female', alpha = 0.9)
plt.plot(sub_new["educ"], lm_income2.params[0] + lm_income2.params[1] * 0 +
lm_income2.params[2] * sub_new["educ"] + lm_income2.params[3] * 0 *
sub_new["educ"], 'r', label = 'male', alpha = 0.9)
plt.title("Men vs women on income with interaction")
plt.xlabel("educ")
plt.ylabel("realrinc")
plt.show()
```

Extra Credit

Try this on the **log** scale and see what difference it makes, if any.

Notes on Interpretation:

- Interactions are multiplicative in nature
- Must always include X1 and X2 if you are including X1*X2
- With interactions included, original Bs for X1 and X2 refer to when X1=0 or when X2=0 … not additive anymore
- Determining statistical significance is trickier

Another example …

Do well-off kids suffer educationally the same amount for each additional sibling, as do non-well-off kids?

A simple multiple regression

```
pd.options.mode.chained_assignment = None

sub_kids = sub[["educ", "sibs", "madeg", "family16", "age"]]
sub_kids["maBA"] = sub_kids['madeg'].isin([3,4])
```

A simple multiple regression

```
lm_maBA = smf.ols(formula = 'educ ~ sibs + maBA' , data = sub_kids).fit()
print (lm_maBA.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   educ   R-squared:                       0.147
Model:                            OLS   Adj. R-squared:                  0.146
Method:                 Least Squares   F-statistic:                     256.1
Date:                Tue, 21 May 2019   Prob (F-statistic):          2.31e-103
Time:                        09:39:57   Log-Likelihood:                -7487.4
No. Observations:                2984   AIC:                         1.498e+04
Df Residuals:                    2981   BIC:                         1.500e+04
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------
Intercept       14.1891      0.090    157.839      0.000        14.013    14.365
maBA[T.True]     2.1164      0.165     12.860      0.000         1.794     2.439
sibs            -0.2859      0.017    -16.493      0.000        -0.320    -0.252
==============================================================================
Omnibus:                      329.893   Durbin-Watson:                   1.778
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              744.276
Skew:                          -0.664   Prob(JB):                    2.41e-162
Kurtosis:                       5.055   Cond. No.                         15.4
==============================================================================

Warnings:
 [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

A simple multiple regression

- For each additional sibling, net of their mom's BA+ degree, a person gets -.286 years less of education (statistically significant)

```
==================================================================================
                    coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Intercept        14.1891      0.090    157.839      0.000      14.013      14.365
maBA[T.True]      2.1164      0.165     12.860      0.000       1.794       2.439
sibs             -0.2859      0.017    -16.493      0.000      -0.320      -0.252
================================================================================
```

A simple multiple regression

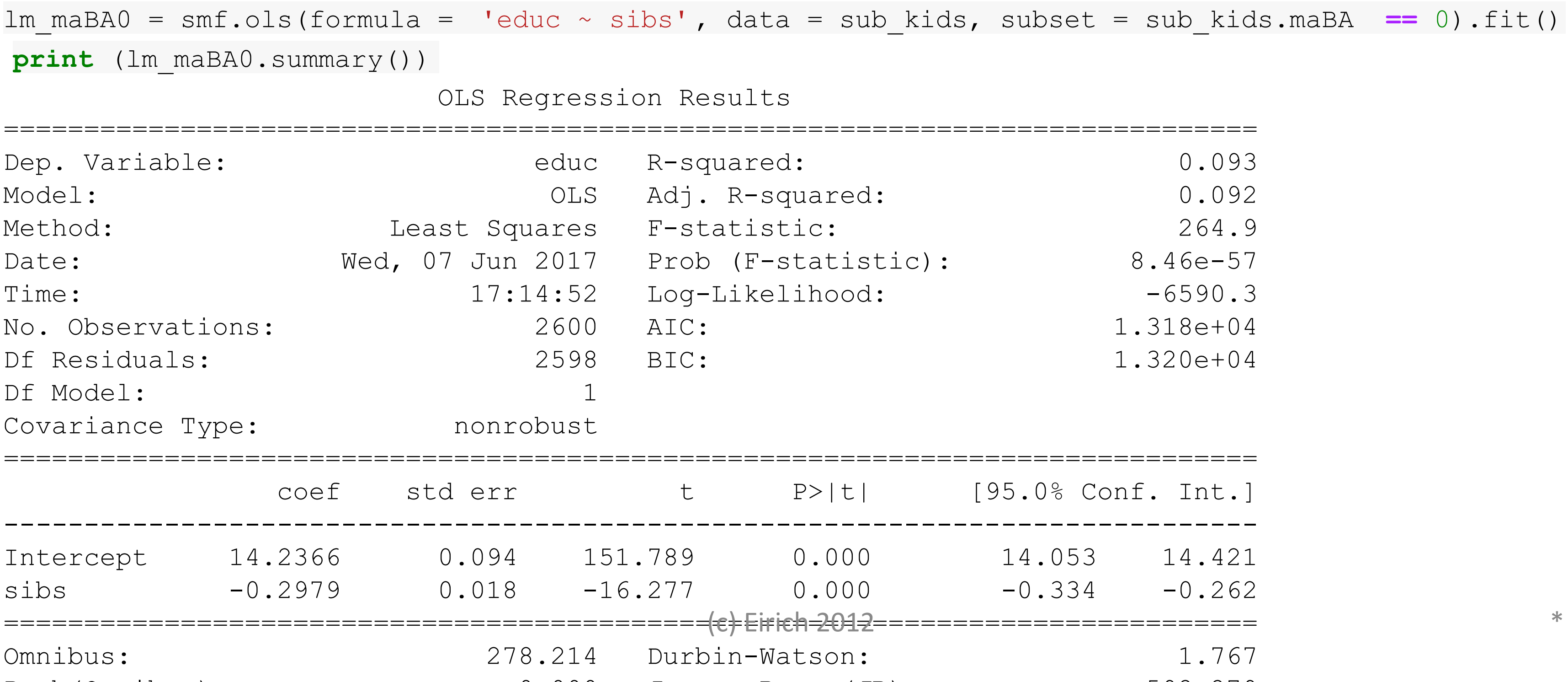
Net of the number of siblings they have, someone’s whose mom has a BA+ degree gets 2.12 years more education than someone’s whose mom has less than a BA (statistically significant)

```
================================================================================
                   coef    std err          t      P>|t|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------
Intercept       14.1891      0.090    157.839      0.000        14.013    14.365
maBA[T.True]     2.1164      0.165     12.860      0.000         1.794     2.439
sibs            -0.2859      0.017    -16.493      0.000        -0.320    -0.252
==============================================================================
```

A simple regression for kids of <BA moms

For kids whose mom has less than a BA, each additional sibling reduces their education by -.298 years of schooling (statistically significant)

(A kid with no sibling (X=0) has 14.24 years of education)



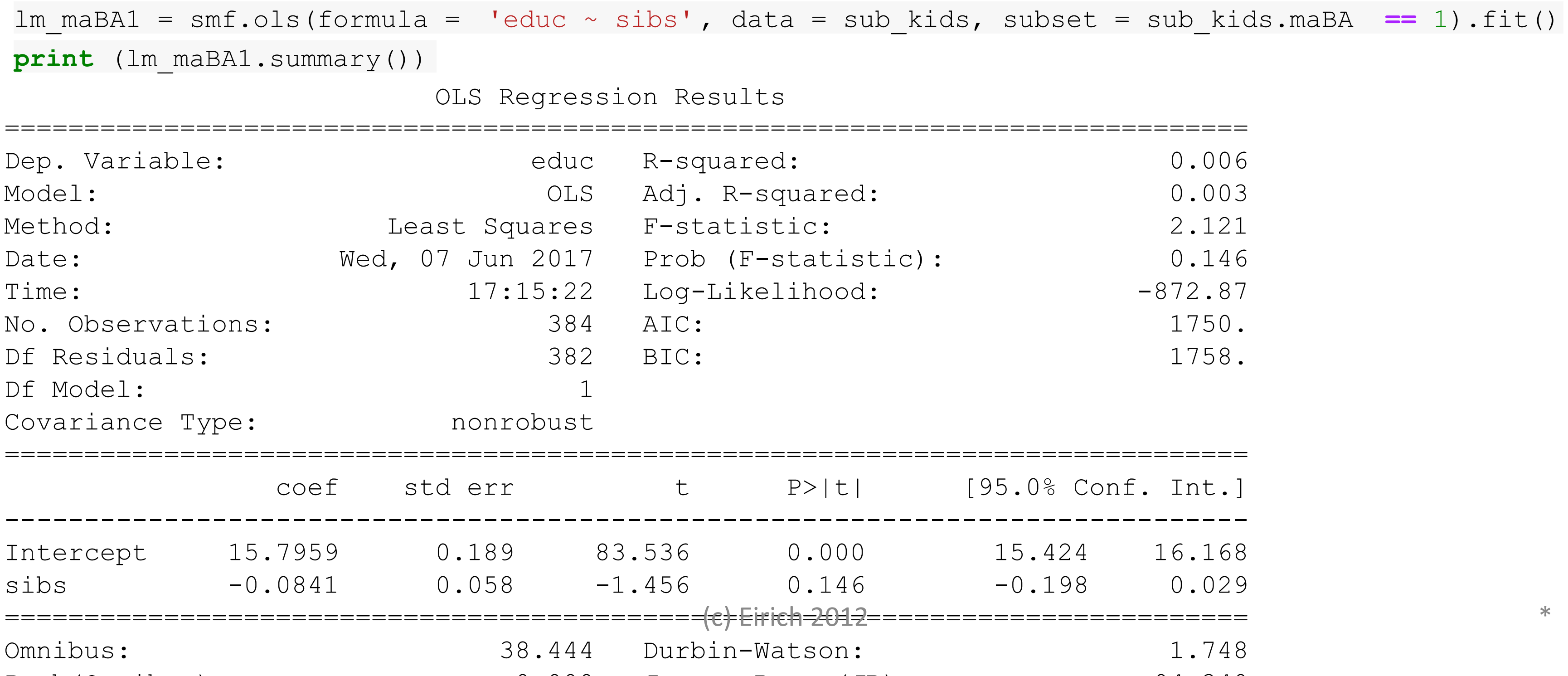
```
lm_maBA0 = smf.ols(formula =  'educ ~ sibs', data = sub_kids, subset = sub_kids.maBA  == 0).fit()
print (lm_maBA0.summary())
```

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                   educ   R-squared:                       0.093
Model:                            OLS   Adj. R-squared:                  0.092
Method:                 Least Squares   F-statistic:                     264.9
Date:                Wed, 07 Jun 2017   Prob (F-statistic):           8.46e-57
Time:                        17:14:52   Log-Likelihood:                -6590.3
No. Observations:                2600   AIC:                         1.318e+04
Df Residuals:                    2598   BIC:                         1.320e+04
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     14.2366      0.094    151.789      0.000        14.053    14.421
sibs          -0.2979      0.018    -16.277      0.000        -0.334    -0.262
==============================================================================
Omnibus:                      278.214   Durbin-Watson:                   1.767
```

A simple regression for kids of BA+ moms

For kids whose mom has a BA+, each additional sibling reduces their education only by -.084 years of schooling (not statistically significant)

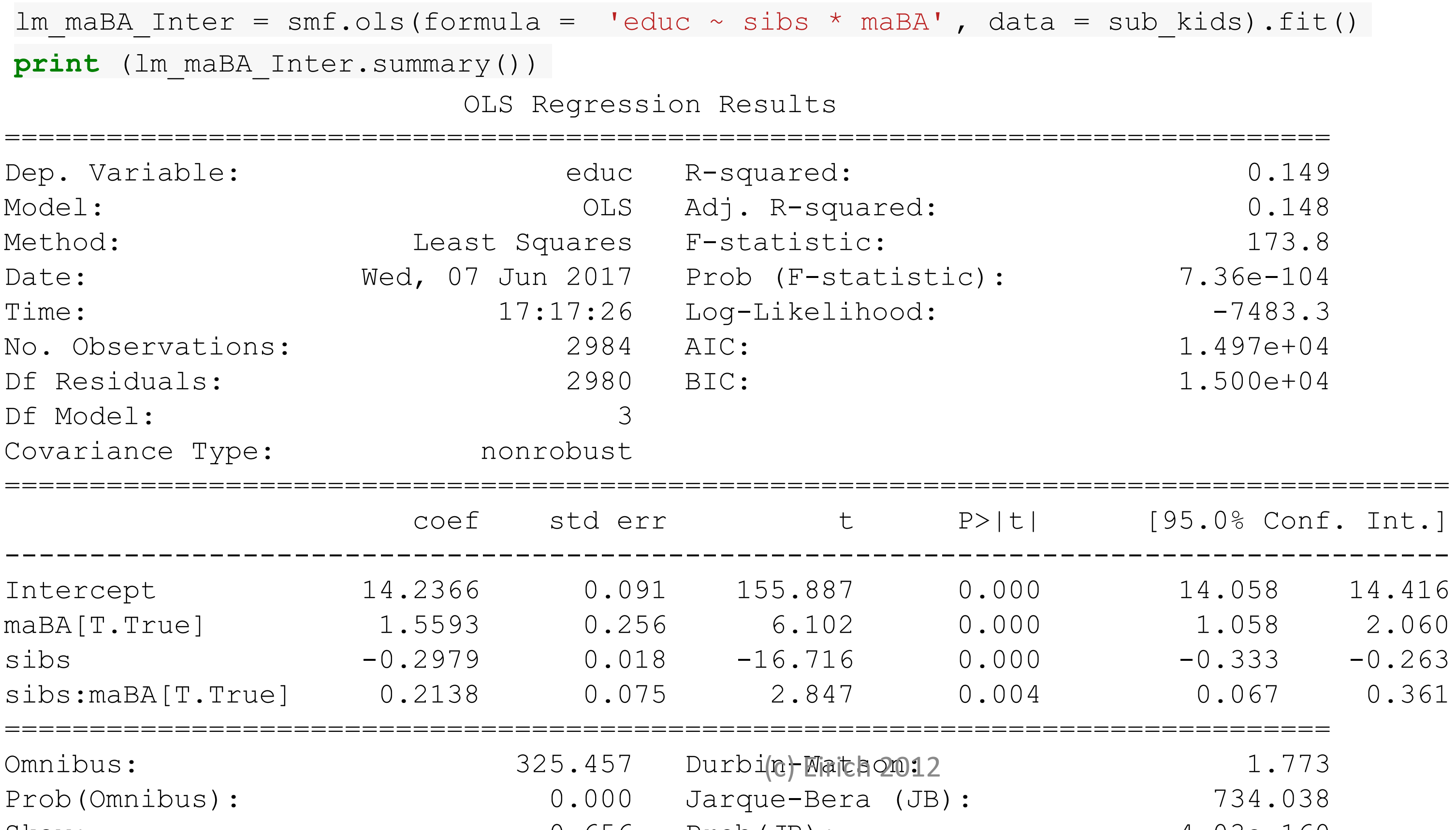
(A kid with no sibling (X=0) has 15.8 years of education)



```
lm_maBA1 = smf.ols(formula =  'educ ~ sibs', data = sub_kids, subset = sub_kids.maBA  == 1).fit()
print (lm_maBA1.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                   educ   R-squared:                       0.006
Model:                            OLS   Adj. R-squared:                  0.003
Method:                 Least Squares   F-statistic:                     2.121
Date:                Wed, 07 Jun 2017   Prob (F-statistic):              0.146
Time:                        17:15:22   Log-Likelihood:                -872.87
No. Observations:                 384   AIC:                             1750.
Df Residuals:                     382   BIC:                             1758.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept     15.7959      0.189     83.536      0.000        15.424    16.168
sibs          -0.0841      0.058     -1.456      0.146        -0.198     0.029
==============================================================================
Omnibus:                       38.444   Durbin-Watson:                   1.748
```

The interaction model I

1. When maBA=0 (i.e., for kids of low educated moms), and with zero siblings, the intercept is the predicted amount of schooling of 14.23



```
lm_maBA_Inter = smf.ols(formula =  'educ ~ sibs * maBA' , data = sub_kids).fit()
print (lm_maBA_Inter.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                   educ   R-squared:                       0.149
Model:                            OLS   Adj. R-squared:                  0.148
Method:                 Least Squares   F-statistic:                     173.8
Date:                Wed, 07 Jun 2017   Prob (F-statistic):          7.36e-104
Time:                        17:17:26   Log-Likelihood:                -7483.3
No. Observations:                2984   AIC:                         1.497e+04
Df Residuals:                    2980   BIC:                         1.500e+04
Df Model:                           3
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------------
Intercept            14.2366      0.091    155.887      0.000        14.058    14.416
maBA[T.True]          1.5593      0.256      6.102      0.000         1.058     2.060
sibs                 -0.2979      0.018    -16.716      0.000        -0.333    -0.263
sibs:maBA[T.True]     0.2138      0.075      2.847      0.004         0.067     0.361
==============================================================================
Omnibus:                      325.457   Durbin-Watson:                   1.773
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              734.038
Skew:                           0.656   Prob(JB):                    4.02e-160
```

The interaction model II

2. When maBA=0 (i.e., for kids of low educated moms), each additional sibling costs a person -.298 years of education.

```
                         coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------------
Intercept             14.2366      0.091    155.887      0.000        14.058    14.416
maBA[T.True]           1.5593      0.256      6.102      0.000         1.058     2.060
sibs                  -0.2979      0.018    -16.716      0.000        -0.333    -0.263
sibs:maBA[T.True]      0.2138      0.075      2.847      0.004         0.067     0.361
==============================================================================
```

The interaction model III

3. When sibs=0, for kids with a BA+ mom, they get 1.56 years more of schooling.

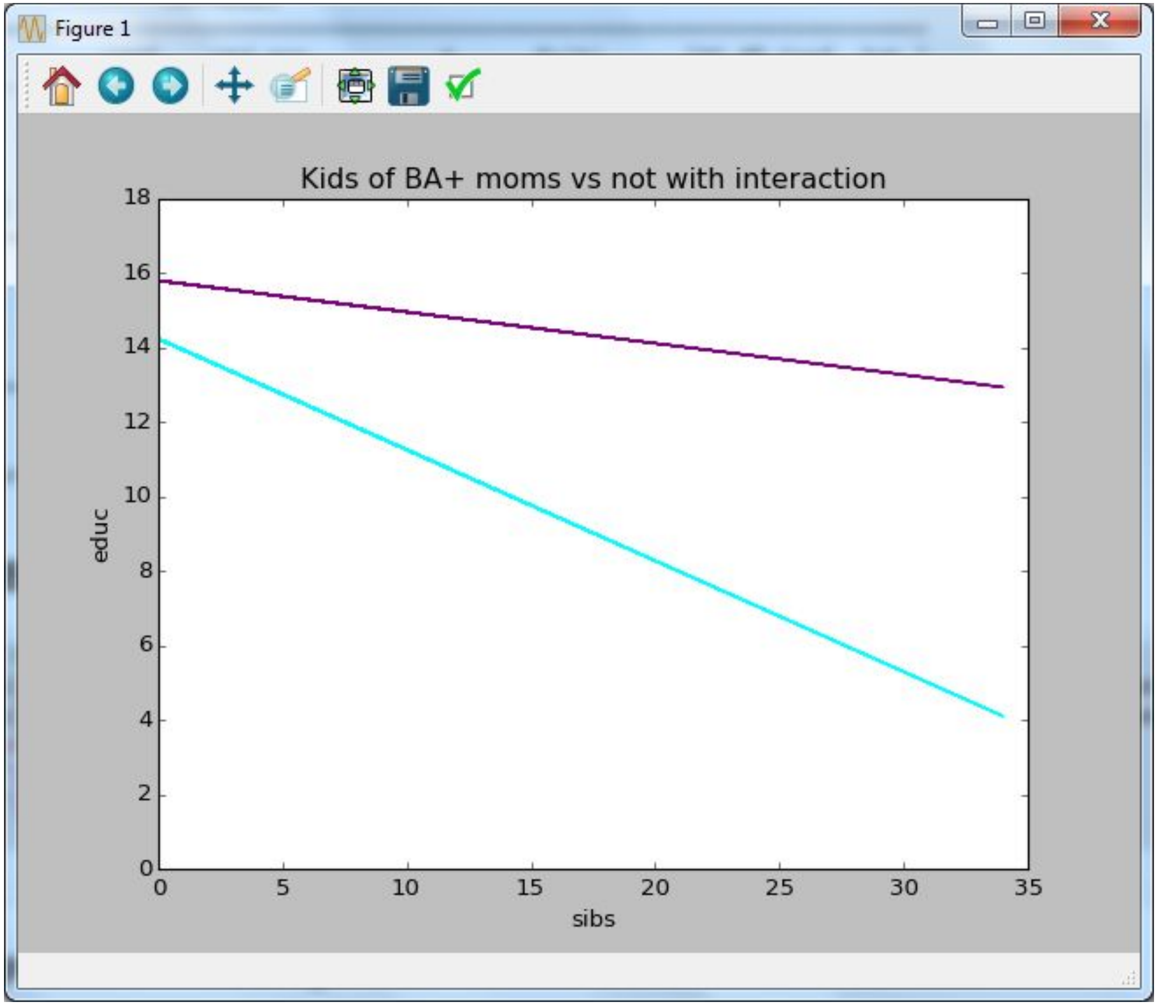
```
                        coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------------
Intercept            14.2366      0.091    155.887      0.000        14.058    14.416
maBA[T.True]          1.5593      0.256      6.102      0.000         1.058     2.060
sibs                 -0.2979      0.018    -16.716      0.000        -0.333    -0.263
sibs:maBA[T.True]     0.2138      0.075      2.847      0.004         0.067     0.361
==============================================================================
```

The interaction model IV

4. For kids of BA+ moms, they gain back .21 for each additional sibling they have. So while they get what kids without BA+ moms get (which is -.298), they add back .21, which means they ultimately get -.08 for each additional sibling

```
                          coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------------
Intercept              14.2366      0.091    155.887      0.000        14.058    14.416
maBA[T.True]            1.5593      0.256      6.102      0.000         1.058     2.060
sibs                   -0.2979      0.018    -16.716      0.000        -0.333    -0.263
sibs:maBA[T.True]       0.2138      0.075      2.847      0.004         0.067     0.361
==============================================================================
```

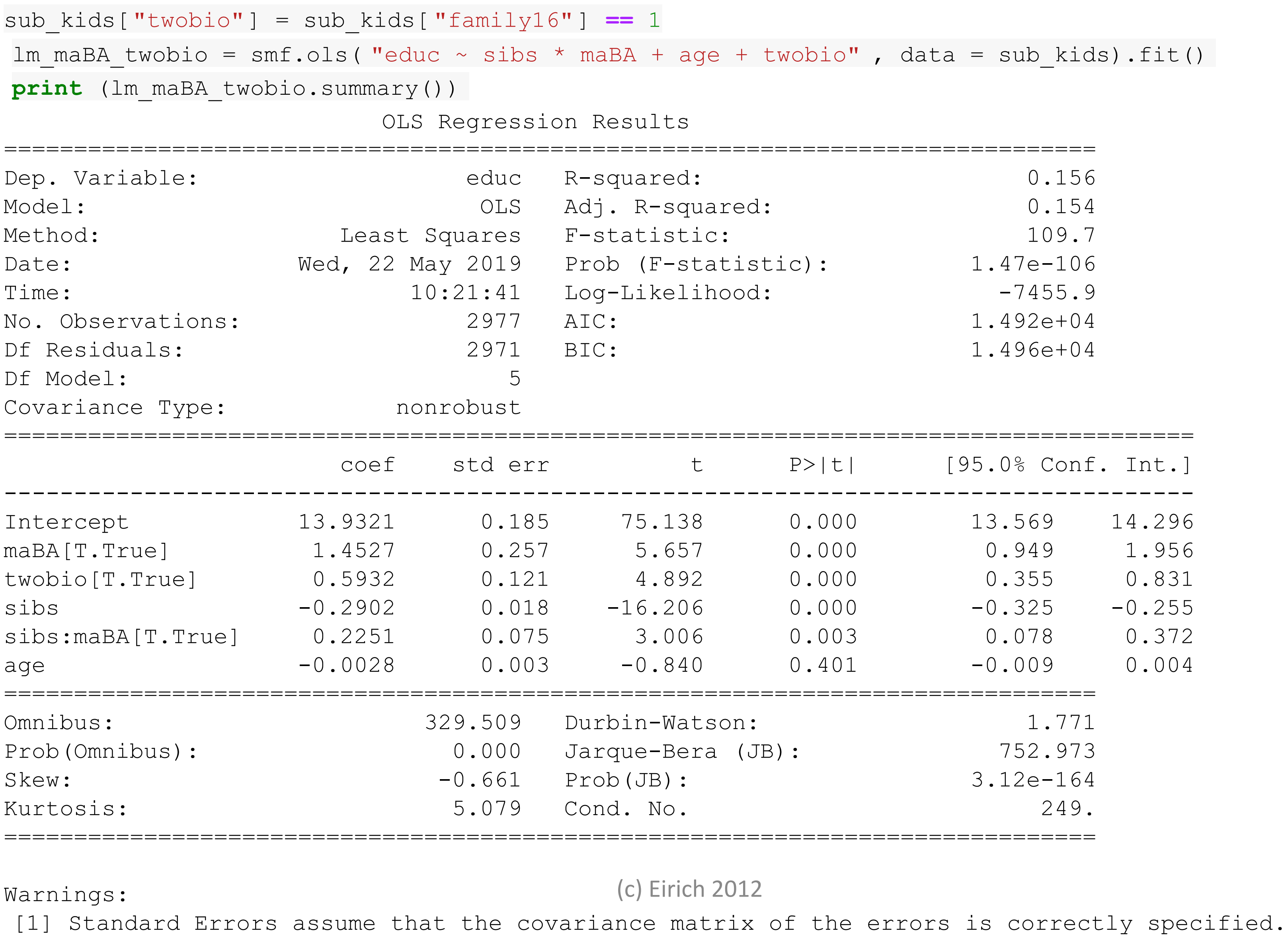
Kids of BA+ moms vs. not, with interaction



Here is that graph

```
plt.axis([0, 35, 0, 18])
plt.plot(sub_kids["sibs"], lm_maBA_Inter.params[0] +
lm_maBA_Inter.params[1] * 0 + lm_maBA_Inter.params[2] *
sub_kids["sibs"] + lm_maBA_Inter.params[3] * 0 * sub_kids["sibs"],
'cyan', label = '<BA', alpha = 0.9)
plt.plot(sub_kids["sibs"], lm_maBA_Inter.params[0] +
lm_maBA_Inter.params[1] * 1 + lm_maBA_Inter.params[2] *
sub_kids["sibs"] + lm_maBA_Inter.params[3] * 1 * sub_kids["sibs"],
'purple', label = 'BA+', alpha = 0.9)
plt.title("Kids of BA+ moms vs not with interaction")
plt.xlabel("sibs")
plt.ylabel("educ")
plt.show()
```

What if we include other variables?



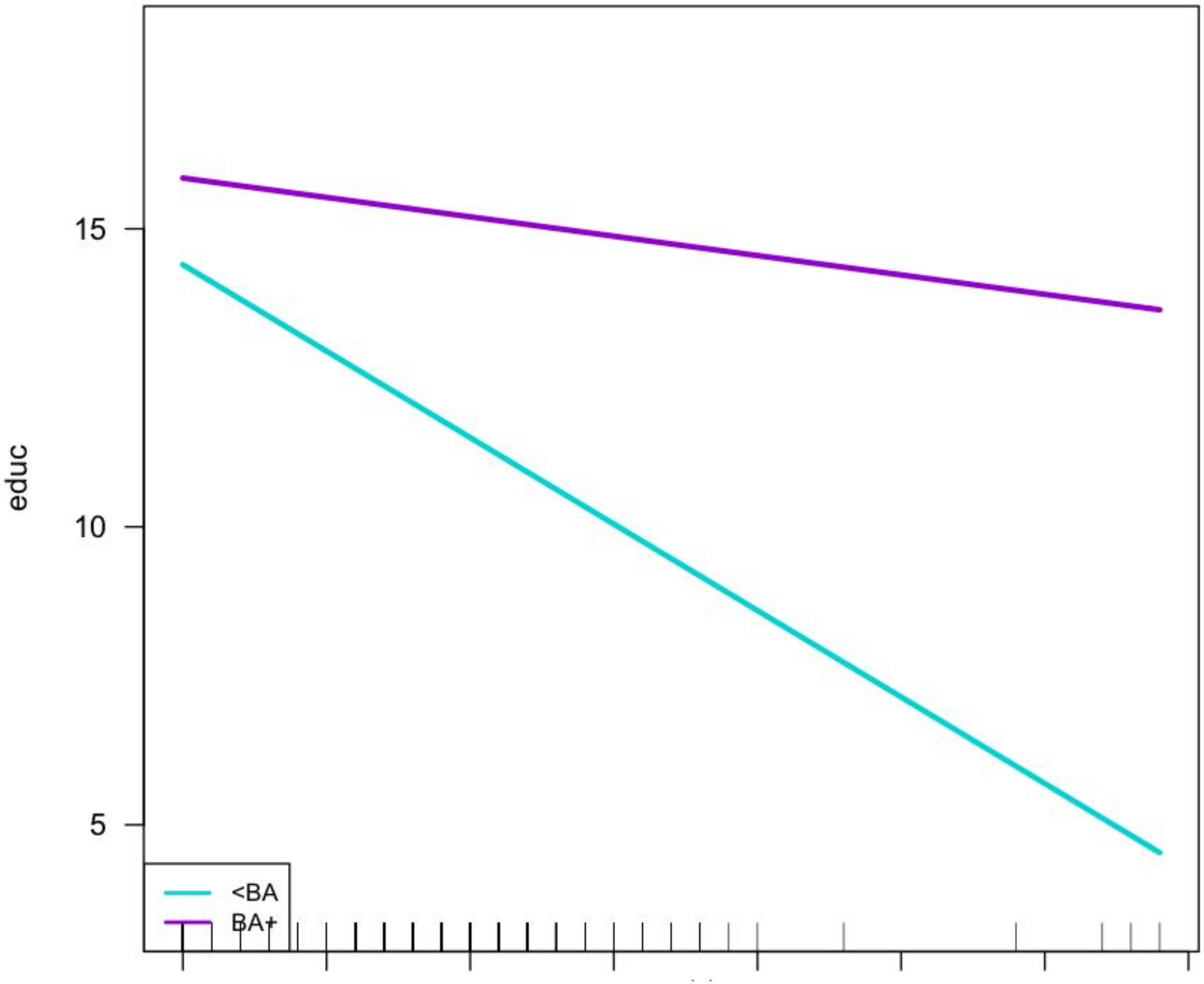
```
sub_kids["twobio"] = sub_kids["family16"] == 1
 lm_maBA_twobio = smf.ols( "educ ~ sibs * maBA + age + twobio" , data = sub_kids).fit()
 print (lm_maBA_twobio.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                   educ   R-squared:                       0.156
Model:                            OLS   Adj. R-squared:                  0.154
Method:                 Least Squares   F-statistic:                     109.7
Date:                Wed, 22 May 2019   Prob (F-statistic):          1.47e-106
Time:                        10:21:41   Log-Likelihood:                -7455.9
No. Observations:                2977   AIC:                         1.492e+04
Df Residuals:                    2971   BIC:                         1.496e+04
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------------
Intercept            13.9321      0.185     75.138      0.000        13.569    14.296
maBA[T.True]          1.4527      0.257      5.657      0.000         0.949     1.956
twobio[T.True]        0.5932      0.121      4.892      0.000         0.355     0.831
sibs                 -0.2902      0.018    -16.206      0.000        -0.325    -0.255
sibs:maBA[T.True]     0.2251      0.075      3.006      0.003         0.078     0.372
age                  -0.0028      0.003     -0.840      0.401        -0.009     0.004
==============================================================================
Omnibus:                      329.509   Durbin-Watson:                   1.771
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              752.973
Skew:                          -0.661   Prob(JB):                    3.12e-164
Kurtosis:                       5.079   Cond. No.                         249.
==============================================================================

Warnings:
 [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

What if we include other variables?

```
=====================================================================================
                           coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------------
Intercept               13.9321      0.185     75.138      0.000        13.569    14.296
maBA[T.True]             1.4527      0.257      5.657      0.000         0.949     1.956
twobio[T.True]           0.5932      0.121      4.892      0.000         0.355     0.831
sibs                    -0.2902      0.018    -16.206      0.000        -0.325    -0.255
sibs:maBA[T.True]        0.2251      0.075      3.006      0.003         0.078     0.372
age                     -0.0028      0.003     -0.840      0.401        -0.009     0.004
==============================================================================
```

Net of other factors, kids (of BA+ moms) gain back .225 for each additional sibling they have – which means they ultimately lose -.075 for each additional sibling

Here is the code for that graph



*

Code here (in R)

```
# Or using visreg

visreg(lm(educ ~ sibs + maBA + sibs:maBA + age + twobio, data = sub),

       xvar = "sibs", by = "maBA", overlay=T, partial = F, band = F, legend =
F,

       line = list(col = c("cyan3", "purple3")))

legend("bottomleft", c("<BA", "BA+"), lwd = 2, col = c("cyan3", "purple3"),
cex = 0.8)
```

Let’s try another one …

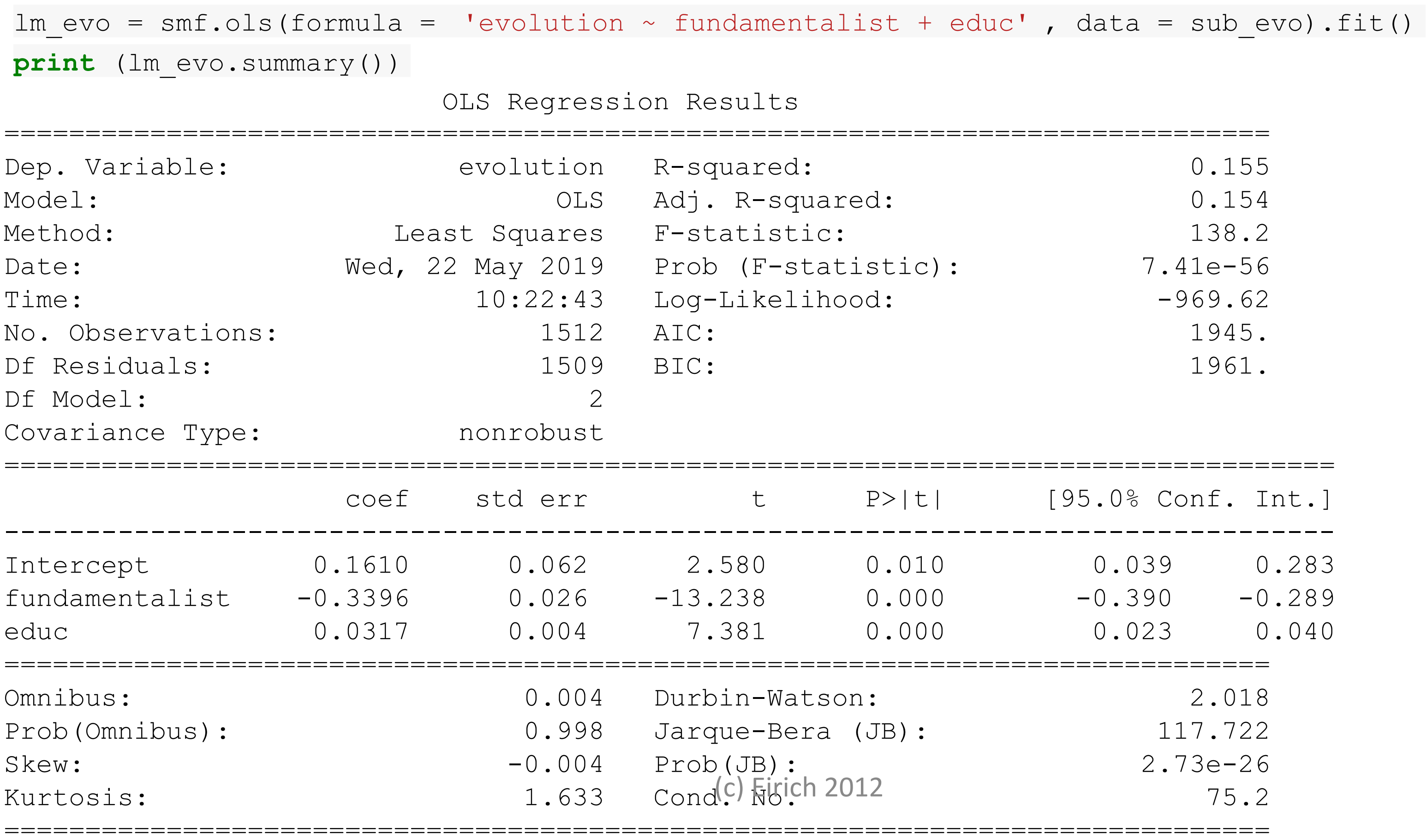
Does education alter fundamentalists opinion on evolution?

The recodes ...

```
pd.options.mode.chained_assignment = None

sub_evo = sub[["educ", "fund", "evolved", "family16", "age"]]
fund_dummy = {1:1, 2:0, 3:0}
sub_evo["fundamentalist"] = sub_evo["fund"].map(fund_dummy.get)
evolved_dummy = {1:1, 2:0}
sub_evo["evolution"] = sub_evo["evolved"].map(evolved_dummy.get)
```

A simple multiple regression

A person who belongs to a fundamentalist religion is -.34 points lower in believing in evolution, net of education.



```
lm_evo = smf.ols(formula =  'evolution ~ fundamentalist + educ' , data = sub_evo).fit()
print (lm_evo.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              evolution   R-squared:                       0.155
Model:                            OLS   Adj. R-squared:                  0.154
Method:                 Least Squares   F-statistic:                     138.2
Date:                Wed, 22 May 2019   Prob (F-statistic):           7.41e-56
Time:                        10:22:43   Log-Likelihood:                -969.62
No. Observations:                1512   AIC:                             1945.
Df Residuals:                    1509   BIC:                             1961.
Df Model:                           2
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
Intercept          0.1610      0.062      2.580      0.010         0.039     0.283
fundamentalist    -0.3396      0.026    -13.238      0.000        -0.390    -0.289
educ               0.0317      0.004      7.381      0.000         0.023     0.040
==============================================================================
Omnibus:                        0.004   Durbin-Watson:                   2.018
Prob(Omnibus):                  0.998   Jarque-Bera (JB):              117.722
Skew:                          -0.004   Prob(JB):                     2.73e-26
Kurtosis:                       1.633   Cond. No.                         75.2
==============================================================================
```

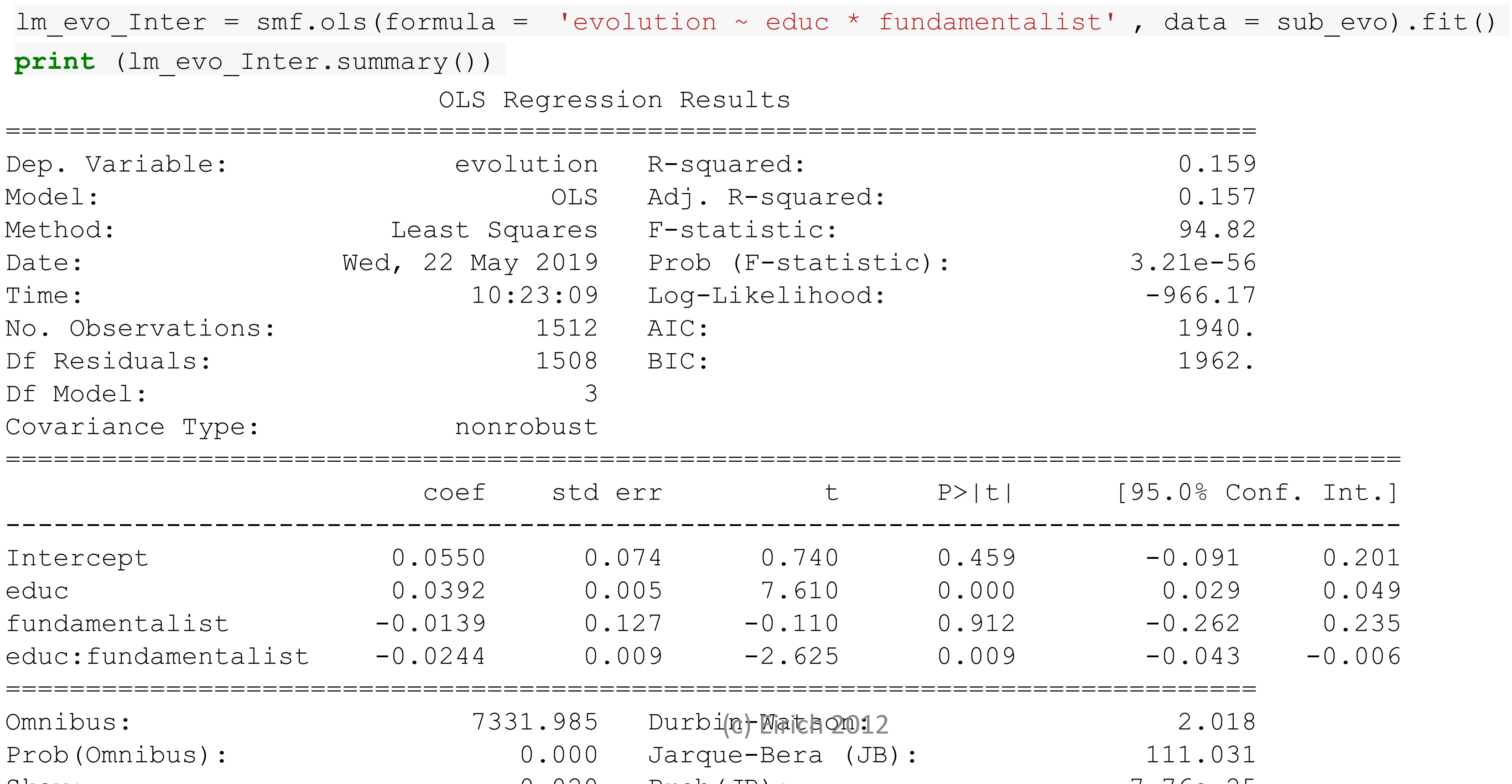
A simple multiple regression

For each year more of education someone has, they have .03 more points of believing in evolution, net of religion.

```
===================================================================================
                     coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------
Intercept          0.1610      0.062      2.580      0.010          0.039     0.283
fundamentalist    -0.3396      0.026    -13.238      0.000         -0.390    -0.289
educ               0.0317      0.004      7.381      0.000          0.023     0.040
===============================================================================
```

The interaction model I

1. When educ=0 and fundamentalist=0 (meaning for a non-fundamentalist), they are predicted to have 0.055 points of believing in evolution



```
lm_evo_Inter = smf.ols(formula =  'evolution ~ educ * fundamentalist' , data = sub_evo).fit()
print (lm_evo_Inter.summary())
                           OLS Regression Results
==============================================================================
Dep. Variable:              evolution   R-squared:                       0.159
Model:                            OLS   Adj. R-squared:                  0.157
Method:                 Least Squares   F-statistic:                     94.82
Date:                Wed, 22 May 2019   Prob (F-statistic):           3.21e-56
Time:                        10:23:09   Log-Likelihood:                -966.17
No. Observations:                1512   AIC:                             1940.
Df Residuals:                    1508   BIC:                             1962.
Df Model:                           3
Covariance Type:            nonrobust
=========================================================================================
                          coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------------
Intercept               0.0550      0.074      0.740      0.459        -0.091     0.201
educ                    0.0392      0.005      7.610      0.000         0.029     0.049
fundamentalist         -0.0139      0.127     -0.110      0.912        -0.262     0.235
educ:fundamentalist    -0.0244      0.009     -2.625      0.009        -0.043    -0.006
==============================================================================
Omnibus:                     7331.985   Durbin-Watson:                   2.018
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              111.031
```

The interaction model II

2. When educ=0, a fundamentalist is -0.0139 (not stat sig.) points lower on believing in evolution than a non-fundamentalist

```
=======================================================================================
                          coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------------
Intercept               0.0550      0.074      0.740      0.459        -0.091     0.201
educ                    0.0392      0.005      7.610      0.000         0.029     0.049
fundamentalist         -0.0139      0.127     -0.110      0.912        -0.262     0.235
educ:fundamentalist    -0.0244      0.009     -2.625      0.009        -0.043    -0.006
===============================================================================
```

The interaction model III

3. When fund=0 (i.e., for a non-fundamentalist), each additional year of education increases someone’s belief in evolution by 0.039 points

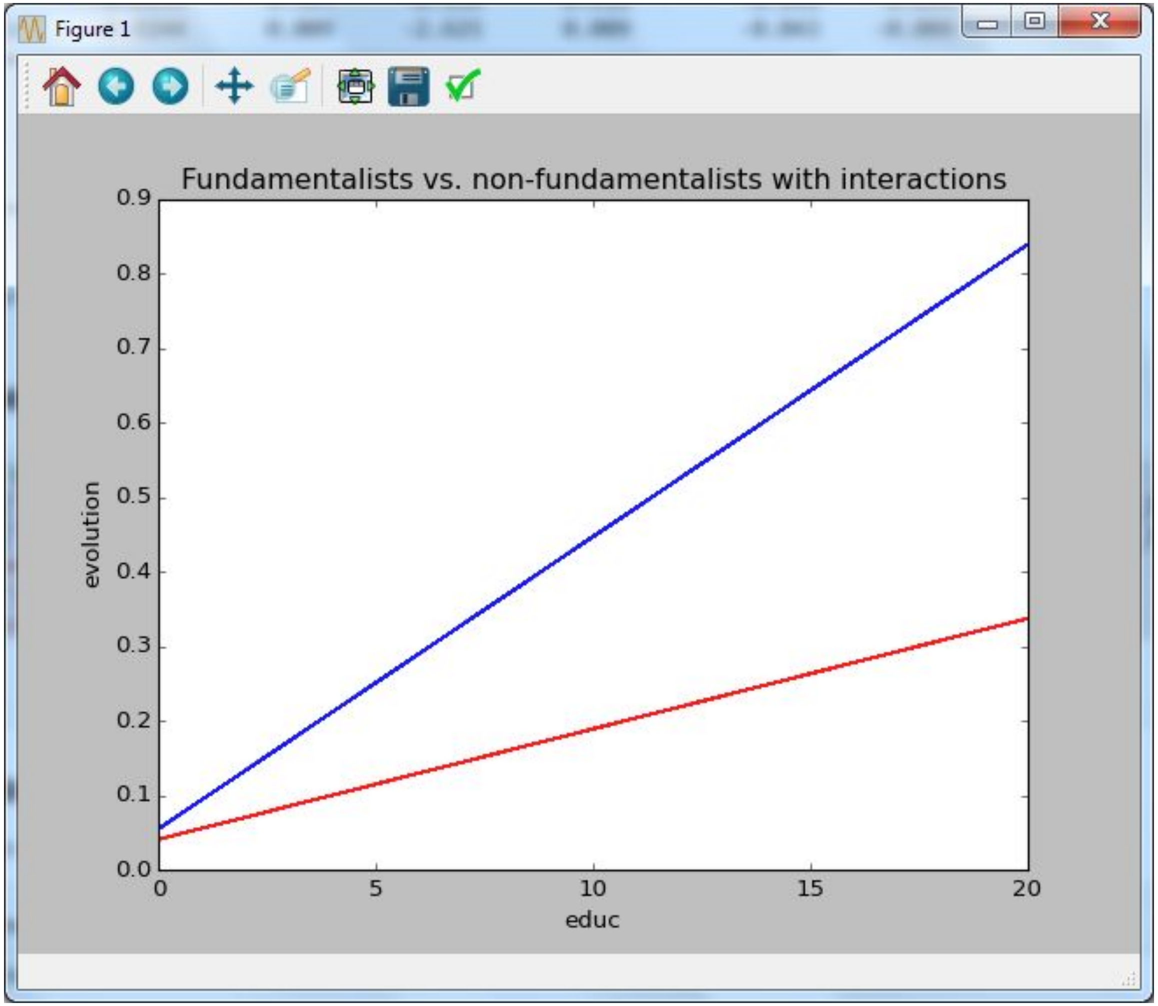
```
=======================================================================================
                          coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------------
Intercept               0.0550      0.074      0.740      0.459        -0.091     0.201
educ                    0.0392      0.005      7.610      0.000         0.029     0.049
fundamentalist         -0.0139      0.127     -0.110      0.912        -0.262     0.235
educ:fundamentalist    -0.0244      0.009     -2.625      0.009        -0.043    -0.006
==============================================================================
```

The interaction model IV

4. Fundamentalists get the 0.039 points on the evolution scale that non-fundamentalists get for each year more of education, but then they lose -0.024 points, for a total of 0.015 points for each year of education for fundamentalists.

```
=======================================================================================
                          coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------------
Intercept               0.0550      0.074      0.740      0.459        -0.091     0.201
educ                    0.0392      0.005      7.610      0.000         0.029     0.049
fundamentalist         -0.0139      0.127     -0.110      0.912        -0.262     0.235
educ:fundamentalist    -0.0244      0.009     -2.625      0.009        -0.043    -0.006
==============================================================================
```

Fundamentalists vs. non-fundamentalists, with interactions



Here is that graph’s code

```
plt.axis([0, 20, 0, 0.9])
plt.plot(sub_evo["educ"], lm_evo_Inter.params[0] + lm_evo_Inter.params[1] *
sub_evo["educ"] + lm_evo_Inter.params[2] * 0 + lm_evo_Inter.params[3] * 0 *
sub_evo["educ"], 'blue', label = 'Not Fundamentalist', alpha = 0.9)
plt.plot(sub_evo["educ"], lm_evo_Inter.params[0] + lm_evo_Inter.params[1] *
sub_evo["educ"] + lm_evo_Inter.params[2] * 1 + lm_evo_Inter.params[3] * 1 *
sub_evo["educ"], 'red', label = 'Fundamentalist', alpha = 0.9)
plt.title("Fundamentalists vs. non-fundamentalists with interactions")
plt.xlabel("educ")
plt.ylabel("evolution")
plt.show()
```

But …

Almost nobody has zero years of education, so why don’t we find a better value to set to zero, like – say – the mean.

We can do that through centering:

```
sub_evo["educ"].mean()

13.293398533007334
```

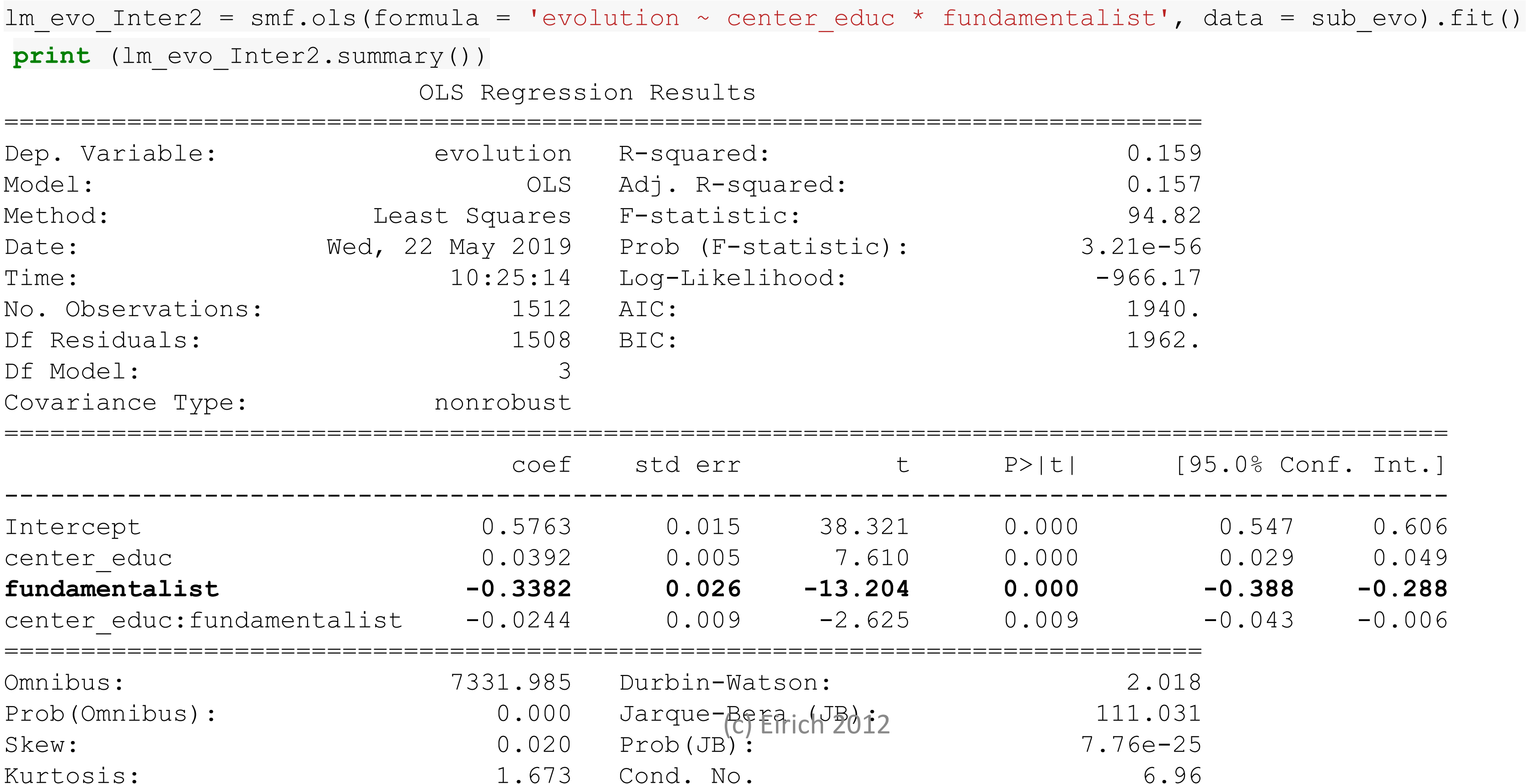
I start with this recode …

```
pd.options.mode.chained_assignment =  None

sub_evo["center_educ"] = sub_evo["educ"] - sub_evo["educ"].mean()

sub_evo["center_educ"].describe().map(lambda x: round(x, 4))
```

```
count    4499.00
mean        0.00
std         3.23
min       -13.29
25%        -1.29
50%        -0.29
75%         2.71
max         6.71
 Name: center_educ, dtype: float64
```

The interaction model I

1. When educ=13.29 (or centereduc=0), a fundamentalist is -.34 points lower on believing in evolution than a non-fundamentalist.



```
lm_evo_Inter2 = smf.ols(formula = 'evolution ~ center_educ * fundamentalist', data = sub_evo).fit()
print (lm_evo_Inter2.summary())
                           OLS Regression Results                            
==============================================================================
Dep. Variable:              evolution   R-squared:                       0.159
Model:                            OLS   Adj. R-squared:                  0.157
Method:                 Least Squares   F-statistic:                     94.82
Date:                Wed, 22 May 2019   Prob (F-statistic):           3.21e-56
Time:                        10:25:14   Log-Likelihood:                -966.17
No. Observations:                1512   AIC:                             1940.
Df Residuals:                    1508   BIC:                             1962.
Df Model:                           3                                         
Covariance Type:            nonrobust                                         
===============================================================================================
                                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------------------
Intercept                      0.5763      0.015     38.321      0.000         0.547     0.606
center_educ                    0.0392      0.005      7.610      0.000         0.029     0.049
fundamentalist                -0.3382      0.026    -13.204      0.000        -0.388    -0.288
center_educ:fundamentalist    -0.0244      0.009     -2.625      0.009        -0.043    -0.006
==============================================================================
Omnibus:                     7331.985   Durbin-Watson:                   2.018
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              111.031
Skew:                           0.020   Prob(JB):                     7.76e-25
Kurtosis:                       1.673   Cond. No.                         6.96
```

Remember the original model …

1. When educ=0, a fundamentalist is -0.014 points lower on believing in evolution than a non-fundamentalist.

```
=======================================================================================
                          coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------------
Intercept               0.0550      0.074      0.740      0.459        -0.091     0.201
educ                    0.0392      0.005      7.610      0.000         0.029     0.049
fundamentalist         -0.0139      0.127     -0.110      0.912        -0.262     0.235
educ:fundamentalist    -0.0244      0.009     -2.625      0.009        -0.043    -0.006
==============================================================================
```

About centering …

Only need to center continuous variables used in the interactions

The statistical significance changes because we are now looking at the change at the mean (where we have lots of data), not at zero (where we had almost no data).

Wait a minute: Why would I think there is an interaction here in the first place?

These previous examples have all been instances of exacerbating (or amplifying) effects

Now let's look at a diminishing (or redundant) effects example

Remember Wordsum & Marriage

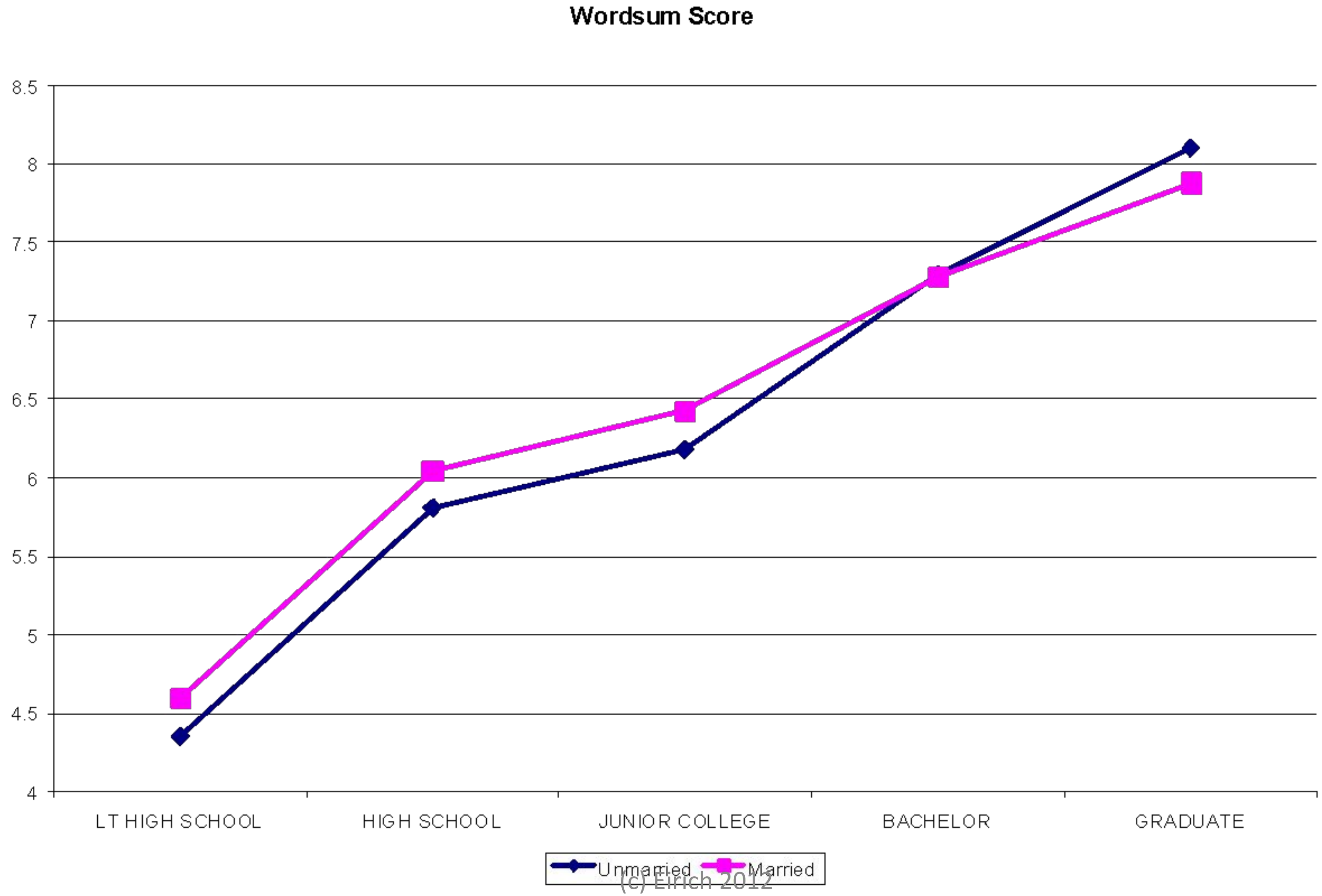
Wordsum, by Marriage & Educ

```
pd.options.mode.chained_assignment = None

sub_word = sub[["marital", "wordsum", "educ", "speduc"]]
sub_word["married"] = sub_word["marital"] == 1
```

Wordsum, by Marriage & Educ

```
lm_wordsum = smf.ols(formula =  'wordsum ~ married + educ' , data = sub_word).fit()
print (lm_wordsum.summary())
                           OLS Regression Results
==============================================================================
Dep. Variable:                wordsum   R-squared:                       0.191
Model:                            OLS   Adj. R-squared:                  0.190
Method:                 Least Squares   F-statistic:                     163.3
Date:                Wed, 22 May 2019   Prob (F-statistic):           2.15e-64
Time:                        10:27:16   Log-Likelihood:                -2785.0
No. Observations:                1388   AIC:                             5576.
Df Residuals:                    1385   BIC:                             5592.
Df Model:                           2
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------
Intercept           2.4666      0.212     11.617      0.000         2.050     2.883
married[T.True]     0.0548      0.097      0.566      0.571        -0.135     0.245
educ                0.2721      0.015     18.024      0.000         0.243     0.302
==============================================================================
Omnibus:                       67.260   Durbin-Watson:                   1.932
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               82.065
Skew:                          -0.496   Prob(JB):                     1.51e-18
Kurtosis:                       3.661   Cond. No.                         61.3
==============================================================================
```

Interaction with 2 Continuous Variables



Wordsum, by Marriage & Educ

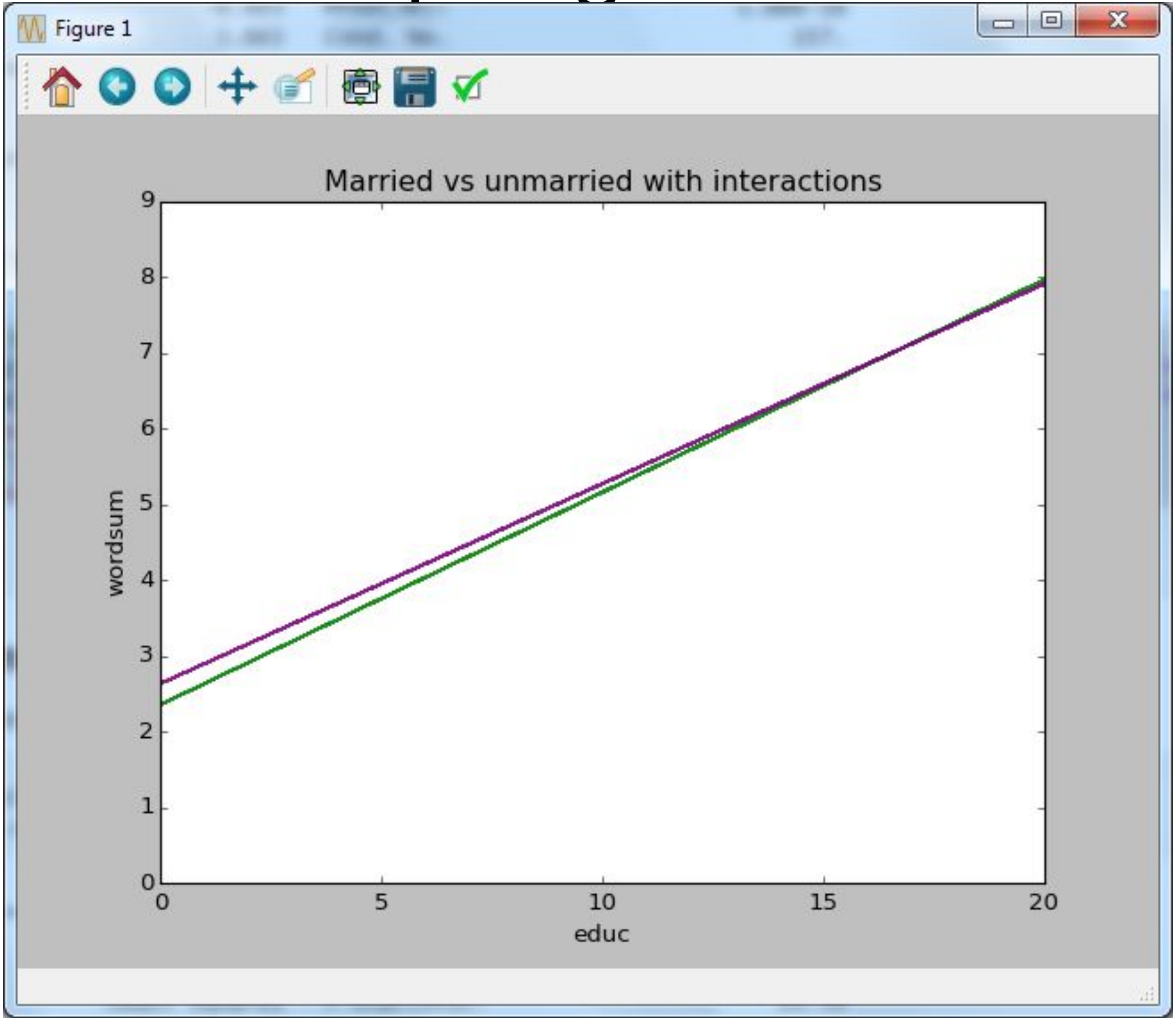
```
lm_wordsum2 = smf.ols(formula =  'wordsum ~ married * educ' , data = sub_word).fit()
print (lm_wordsum2.summary())
                           OLS Regression Results
==============================================================================
Dep. Variable:                wordsum   R-squared:                       0.191
Model:                            OLS   Adj. R-squared:                  0.189
Method:                 Least Squares   F-statistic:                     108.9
Date:                Wed, 22 May 2019   Prob (F-statistic):           2.67e-63
Time:                        10:28:20   Log-Likelihood:                -2784.9
No. Observations:                1388   AIC:                             5578.
Df Residuals:                    1384   BIC:                             5599.
Df Model:                           3
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------------
Intercept                2.3537      0.295      7.985      0.000         1.775     2.932
married[T.True]          0.2794      0.418      0.669      0.504        -0.540     1.099
educ                     0.2806      0.022     13.035      0.000         0.238     0.323
married[T.True]:educ    -0.0167      0.030     -0.553      0.581        -0.076     0.043
==============================================================================
Omnibus:                       66.856   Durbin-Watson:                   1.930
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               81.633
Skew:                          -0.493   Prob(JB):                     1.88e-18
Kurtosis:                       3.663   Cond. No.                         157.
==============================================================================

Warnings:
 [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Wordsum, by Marriage & Educ

```
==============================================================================
                          coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept               2.3537      0.295      7.985      0.000         1.775     2.932
married[T.True]         0.2794      0.418      0.669      0.504        -0.540     1.099
educ                    0.2806      0.022     13.035      0.000         0.238     0.323
married[T.True]:educ   -0.0167      0.030     -0.553      0.581        -0.076     0.043
==============================================================================
```

For married people, they lose -0.0167 (not statistically significant) Wordsum points for each year more educated they are, relative to non-married people

Graphing this relationship



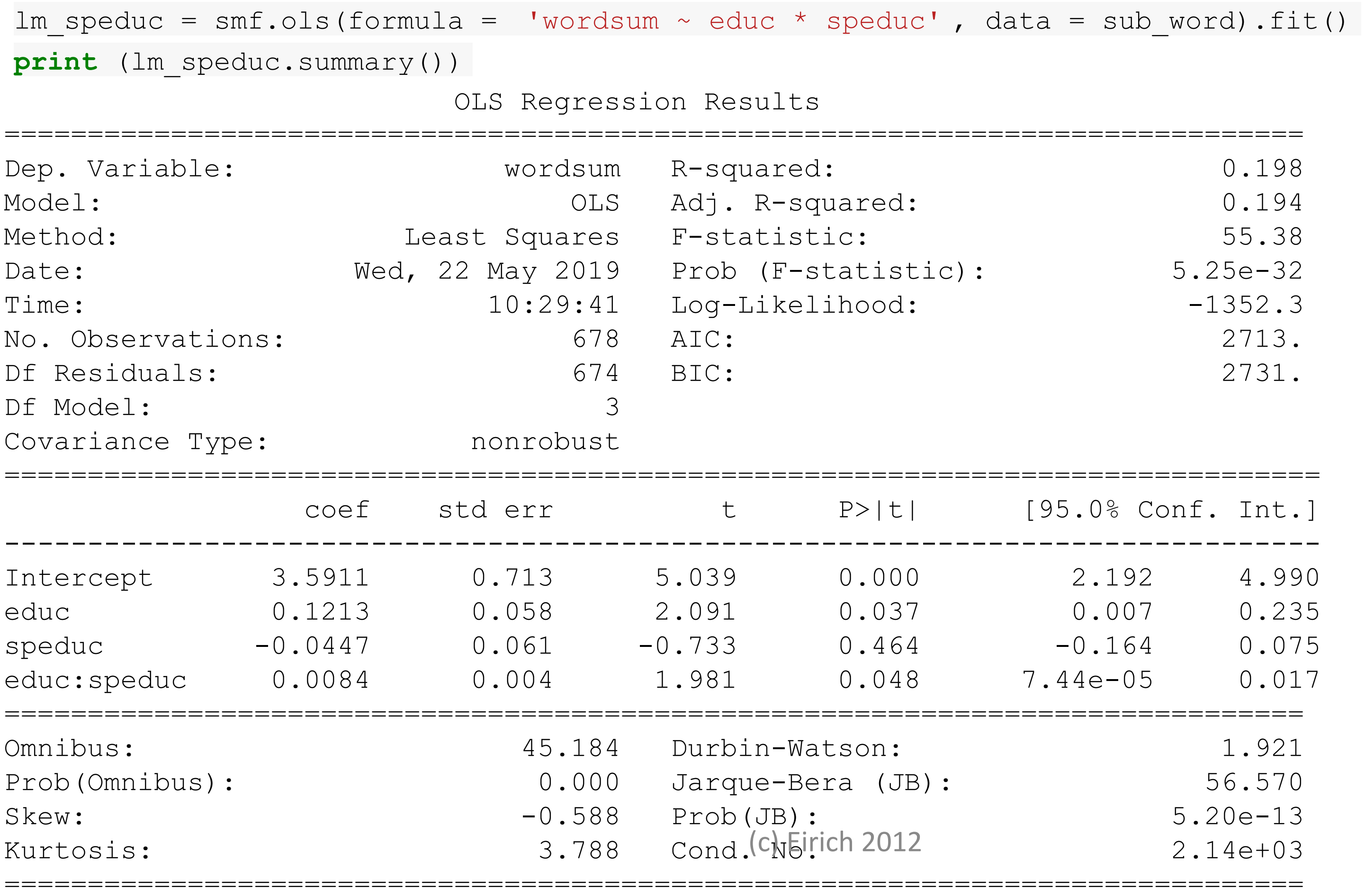
*

Here is that graph’s code

```
plt.axis([0, 20, 0, 9])
plt.plot(sub word["educ"], lm wordsum2.params[0] + lm wordsum2.params[1] * 0
+ lm wordsum2.params[2] * sub word["educ"] + lm wordsum2.params[3] * 0 *
sub word["educ"], 'green', label = 'Unmarried', alpha = 0.9)
plt.plot(sub word["educ"], lm wordsum2.params[0] + lm wordsum2.params[1] * 1
+ lm wordsum2.params[2] * sub word["educ"] + lm wordsum2.params[3] * 1 *
sub word["educ"], 'purple', label = 'Married', alpha = 0.9)
plt.title("Married vs unmarried with interactions")
plt.xlabel("educ")
plt.ylabel("wordsum")
plt.show()
```

A mechanism by which marriage can increase WordSum?

Wordsum, by My Educ & My Spouse’s Educ



```
lm_speduc = smf.ols(formula =  'wordsum ~ educ * speduc' , data = sub_word).fit()
print (lm_speduc.summary())
                           OLS Regression Results
==============================================================================
Dep. Variable:                wordsum   R-squared:                       0.198
Model:                            OLS   Adj. R-squared:                  0.194
Method:                 Least Squares   F-statistic:                     55.38
Date:                Wed, 22 May 2019   Prob (F-statistic):           5.25e-32
Time:                        10:29:41   Log-Likelihood:                -1352.3
No. Observations:                 678   AIC:                             2713.
Df Residuals:                     674   BIC:                             2731.
Df Model:                           3
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------
Intercept       3.5911      0.713      5.039      0.000         2.192     4.990
educ            0.1213      0.058      2.091      0.037         0.007     0.235
speduc         -0.0447      0.061     -0.733      0.464        -0.164     0.075
educ:speduc     0.0084      0.004      1.981      0.048      7.44e-05     0.017
==============================================================================
Omnibus:                       45.184   Durbin-Watson:                   1.921
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               56.570
Skew:                          -0.588   Prob(JB):                     5.20e-13
Kurtosis:                       3.788   Cond. No.                     2.14e+03
==============================================================================
```

Wordsum, by My Educ & My Spouse's Educ

```
===============================================================================
                  coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------
Intercept       3.5911      0.713      5.039      0.000         2.192     4.990
educ            0.1213      0.058      2.091      0.037         0.007     0.235
speduc         -0.0447      0.061     -0.733      0.464        -0.164     0.075
educ:speduc     0.0084      0.004      1.981      0.048      7.44e-05     0.017
==============================================================================
```

The slope on the interaction of R's education and Spouse's education is positive, meaning that it leads to a widening gap as both educ and speduc grow larger

Interpreting continuous by continuous interactions

Wordsum = 3.59 + 0.12*Educ - 0.044*SpEduc + 0.008*Educ*SpEduc

Set SpEduc=**0**, then:

Wordsum = 3.59 + 0.12*Educ + 0.044*(0) + 0.008*Educ*(0)

Wordsum = 3.59 + 0.12*Educ

Interpreting continuous by continuous interactions

If you plug in values for X2, then you can figure out both the intercept and slope for each line …

Interpreting continuous by continuous interactions

Set SpEduc=**10**, then:

Wordsum = 3.59 + 0.12*Educ - 0.044*(10) + 0.008*Educ*(10)

Wordsum = 3.59 + 0.12*Educ - 0.44 + 0.08*Educ

Wordsum = 3.15 + 0.20*Educ

Interpreting continuous by continuous interactions

Set SpEduc=**20**, then:

Wordsum = 3.59 + 0.12*Educ - 0.044*(20) + 0.008*Educ*(20)

Wordsum = 3.59 + 0.12*Educ - 0.88 + 0.16*Educ

Wordsum = 2.71 + 0.28*Educ

Interpreting continuous by continuous interactions

When SpEduc=0, Wordsum = 3.59 + 0.12*Educ

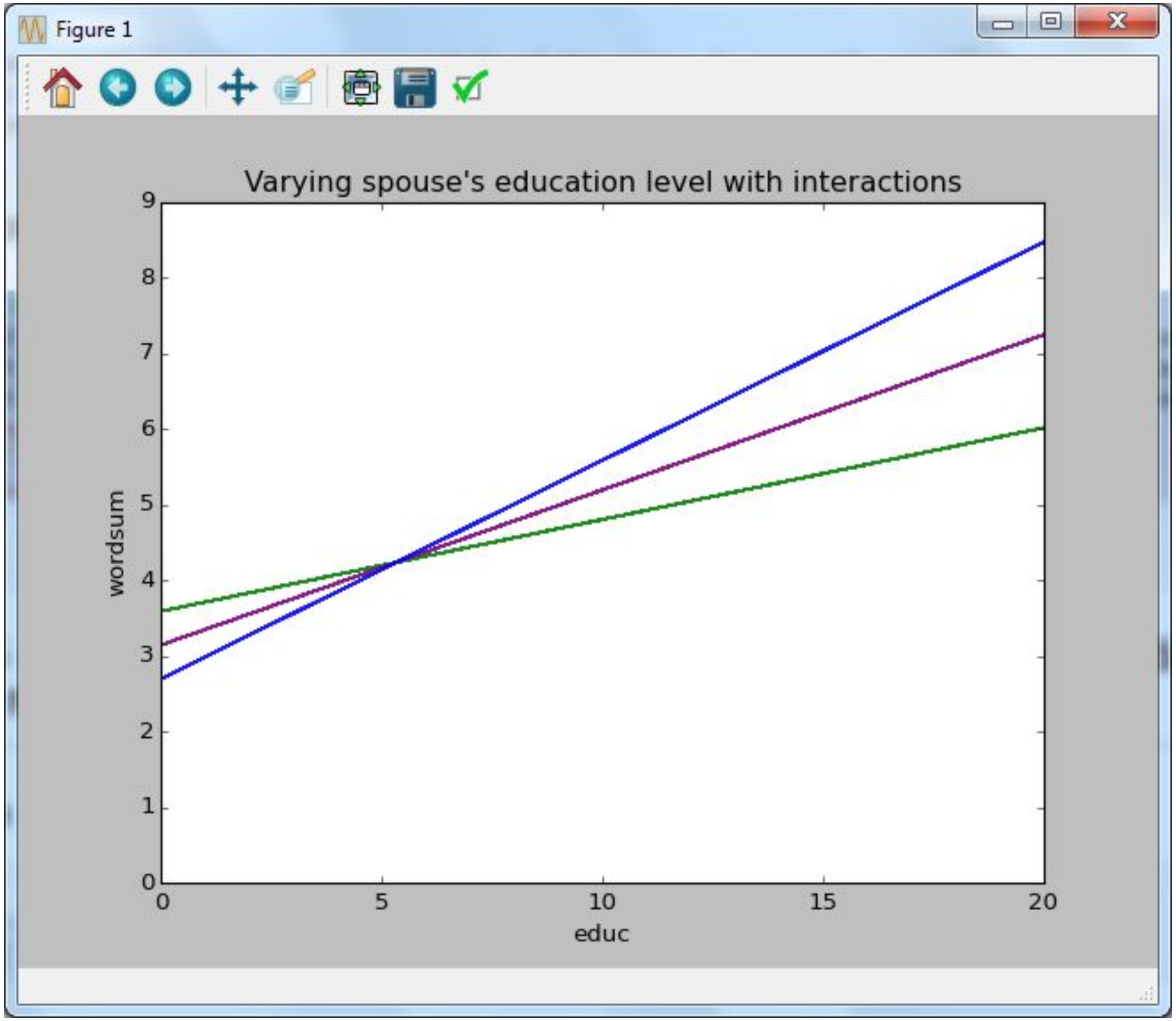
When SpEduc=10, Wordsum = 3.15 + 0.20*Educ

When SpEduc=20, Wordsum = 2.71 + 0.28*Educ

As SpEduc increases, the intercept decreases

As SpEduc increases, the slope increases too

Graphing this relationship



The code

```
plt.axis([0, 20, 0, 9])
plt.plot(sub word["educ"], lm speduc.params[0] + lm speduc.params[1] *
sub word["educ"] + lm speduc.params[2] * 0 + lm speduc.params[3] * 0 *
sub word["educ"], 'green', label = 'SpEduc = 0', alpha = 0.9)
plt.plot(sub word["educ"], lm speduc.params[0] + lm speduc.params[1] *
sub word["educ"] + lm speduc.params[2] * 10 + lm speduc.params[3] * 10 *
sub word["educ"], 'purple', label = 'SpEduc = 10', alpha = 0.9)
plt.plot(sub word["educ"], lm speduc.params[0] + lm speduc.params[1] *
sub word["educ"] + lm speduc.params[2] * 20 + lm speduc.params[3] * 20 *
sub word["educ"], 'blue', label = 'SpEduc = 20', alpha = 0.9)
plt.title("Varying spouse's education level with interactions")
plt.xlabel("educ")
plt.ylabel("wordsum")
plt.show()
```

Another example of an interaction (in STATA, sorry)

Being more educated is associated with improved health. Going to religious services more is associated with improved health. Does someone get even more out of their education and attendance when they are high on both?

Simple regression, in STATA

```
. vreverse health, gen(rhealth)

. reg rhealth educ attend age, beta

      Source |       SS       df       MS              Number of obs =   40522
-------------+------------------------------           F(  3, 40518) = 2168.31
       Model |  4027.32716     3  1342.44239           Prob > F      =  0.0000
    Residual |  25085.5252 40518   .61912052           R-squared     =  0.1383
-------------+------------------------------           Adj R-squared =  0.1383
       Total |  29112.8524 40521  .718463325           Root MSE      =  .78684

------------------------------------------------------------------------------
     rhealth |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        educ |   .0673073   .0012567    53.56   0.000                 .2533878
      attend |   .0237323   .0014629    16.22   0.000                 .0756068
         age |  -.0106901   .0002324   -46.00   0.000                -.2197657
       _cons |   2.549936   .0215238   118.47   0.000                        .
------------------------------------------------------------------------------
```

Both higher education and higher religious attendance are positively predictive of health, net of age

Interaction model

```
. reg rhealth c.educ##c.attend age

      Source |       SS       df       MS              Number of obs =   40522
-------------+------------------------------           F(  4, 40517) = 1631.26
       Model |  4038.14542     4  1009.53635           Prob > F      =  0.0000
    Residual |   25074.707 40517  .618868795           R-squared     =  0.1387
-------------+------------------------------           Adj R-squared =  0.1386
       Total |  29112.8524 40521  .718463325           Root MSE      =  .78668

------------------------------------------------------------------------------
     rhealth |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0747059   .0021703    34.42   0.000     .0704522    .0789597
      attend |   .0473712     .00584     8.11   0.000     .0359247    .0588177
      |
      c.educ#|
    c.attend |  -.0018764   .0004488    -4.18   0.000     -.002756   -.0009967
      |
         age |  -.0107206   .0002325   -46.12   0.000    -.0111762    -.010265
       _cons |   2.458203   .0307323    79.99   0.000     2.397967    2.518439
------------------------------------------------------------------------------
```

As both education and attendance increase together, they have a diminishing effect on someone’s health (note: health was reverse coded), net of age

Interpreting continuous by continuous interactions

Health = 2.45 + 0.074*Educ + 0.047*Attend - 0.0018*Educ*Attend (notice: I don't need to include age, because we can set that to anything constant for each line)

Set Attend=**0**, then:

Health = 2.45 + 0.074*Educ + 0.047*(0) - 0.0018*Educ*(0)

Health = 2.45 + 0.074*Educ

Interpreting continuous by continuous interactions

Health = 2.45 + 0.074*Educ + 0.047*Attend - 0.0018*Educ*Attend

Set Attend=**4**, then:

Health = 2.45 + 0.074*Educ + 0.047*(4) - 0.0018*Educ*(4)

Health = 2.65 + 0.066*Educ

Interpreting continuous by continuous interactions

$$\text{Health} = 2.45 + 0.074*\text{Educ} + 0.047*\text{Attend} - 0.0018*\text{Educ}*\text{Attend}$$

Set Attend=**8**, then:

$$\text{Health} = 2.45 + 0.074*\text{Educ} + 0.047*(8) - 0.0018*\text{Educ}*(8)$$

$$\text{Health} = 2.85 + 0.058*\text{Educ}$$

Interpreting continuous by continuous interactions

When Attend=0, Attend = 2.45 + 0.074*Educ

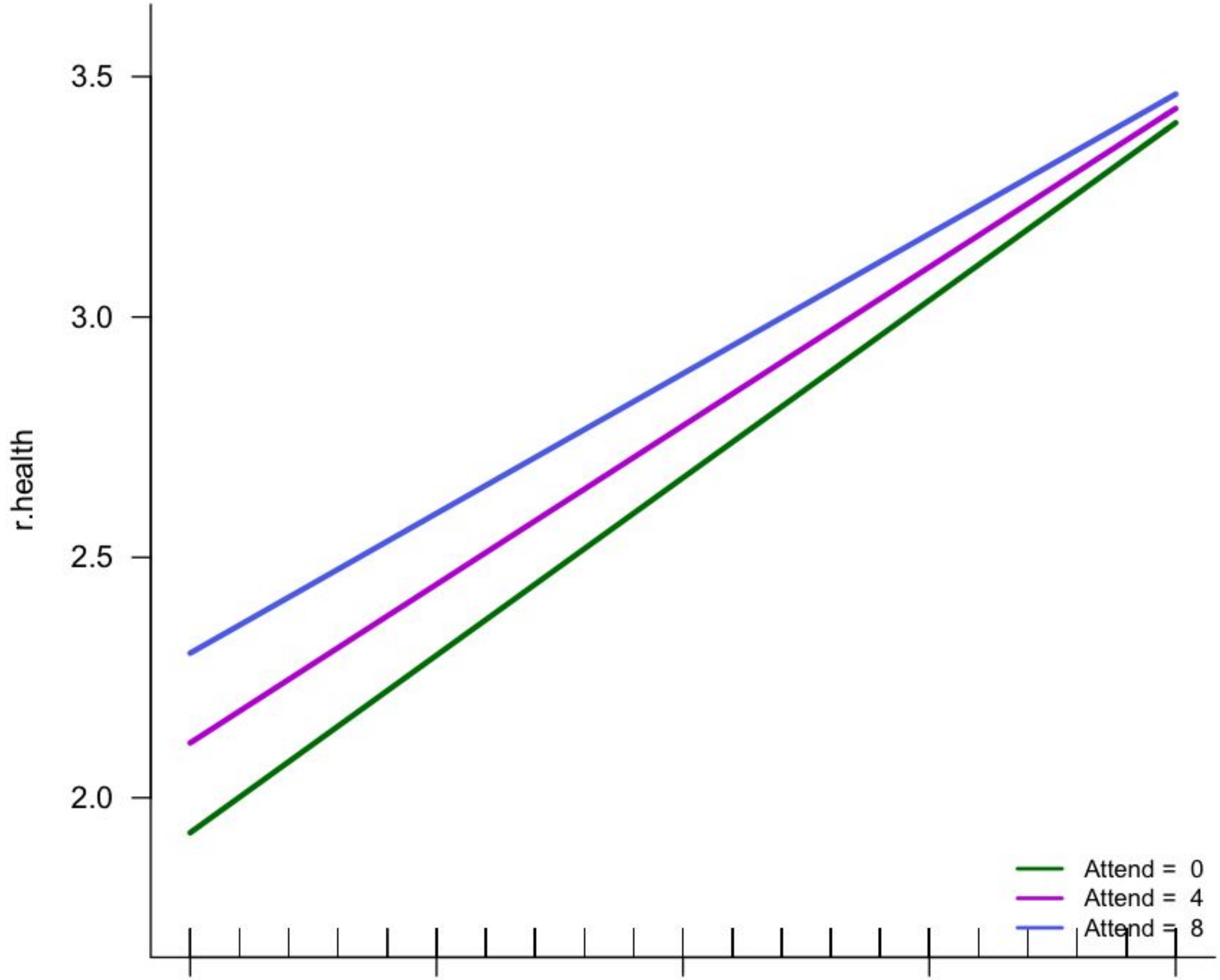
When Attend=4, Health = 2.65 + 0.066*Educ

When Attend=8, Health = 2.85 + 0.058*Educ

As Attend increases, the intercept increases too

But as Attend increases, the slope decreases

Here it is graphed



*

Here is that graph's code (in R)

```
lm.health <- lm(r.health ~ educ*attend + age, data = sub)
summary(lm.health)


# Plotting the relationship (using visreg for practice)
visreg(lm.health, "educ", by = "attend", breaks = c(0,4,8),
      overlay=T, band = F, partial = F, bty = "l", legend = F,
      line = list(col = c("darkgreen", "darkorchid","royalblue")))
legend("bottomright", paste("Attend = ", c(0,4,8)), bty = "n", lwd = 2,
      col = c("darkgreen", "darkorchid","royalblue"), cex = 0.8)
```