In [79]:
```
import numpy as np
import pandas as pd
import os
import matplotlib.pyplot as plt
import statsmodels.formula.api as smf
import seaborn as sn
import math
```

In [80]:
```
os.chdir("C:/Users/jcinterrante/OneDrive - The University of Chicago/Classes/Machine Le

acs_data = pd.read_csv("usa_00002.csv")
acs_data = acs_data[acs_data["INCWAGE"] != 999999]
crosswalk = pd.read_csv("educ crosswalk.csv")
```

2.a.

Use the educd variable to create a continuous measure of education called educdc using the crosswalk

In [81]:
```
acs_data['EDUCD'].unique()

acs_data = acs_data.merge(crosswalk, left_on="EDUCD", right_on="educd")
```

2.b.

Create dummy variables hsdip, coldip, white, black, hispanic, married, female, vet

In [82]:
```
acs_data["hsdip"] = np.where(np.isin(acs_data["EDUCD"], range(62, 84))|np.isin(acs_data
acs_data["coldip"] = np.where(np.isin(acs_data["EDUCD"], range(101, 117)), 1, 0)
acs_data["white"] = np.where(acs_data["RACE"] == 1, 1, 0)
acs_data["black"] = np.where(acs_data["RACE"] == 2, 1, 0)
acs_data["hispanic"] = np.where(np.isin(acs_data["HISPAN"], 1, 4), 1, 0)
acs_data["married"] = np.where(np.isin(acs_data["MARST"], 1|2), 1, 0)
acs_data["female"] = np.where(acs_data["SEX"] == 2, 1, 0)
acs_data["vet"] = np.where(acs_data["VETSTAT"] == 2, 1, 0)
```

2.c.

Create an interaction between each of the education dummy variables (A-B) and education.

In [83]:
```
acs_data["educ_x_hs"] = acs_data["educdc"] * acs_data["hsdip"]
acs_data["educ_x_col"] = acs_data["educdc"] * acs_data["coldip"]
```

2.d.

Create the following: Age squared. The natural log of incwage.

In [84]:
```
acs_data["age_squared"] = np.power(acs_data["AGE"], 2)
```

```
acs_data = acs_data[acs_data['INCWAGE'] > 0]
acs_data["incwage_log"] = np.log(acs_data["INCWAGE"])
```

4.1.

Compute descriptive (summary) statistics for the following variables: year, incwage, lnincwage, educdc, female, age, age2, white, black, hispanic, married, nchild, vet, hsdip, coldip, and the interaction terms. In other words, compute sample means, standard deviations, etc.

In [85]:
```
summary_cols = ["YEAR", "EDUC", "educdc", "female", "AGE", "age_squared",
                "white", "black", "hispanic", "married", "NCHILD","vet","hsdip",
                "coldip", "educ_x_hs", "educ_x_col"]
summary = acs_data.describe(include="all")[summary_cols]
summary
```

Out[85]:

	YEAR	EDUC	educdc	female	AGE	age_squared	white	bla
count	9293.0	9293.000000	9293.000000	9293.000000	9293.000000	9293.000000	9293.000000	9293.0000
mean	2019.0	7.777359	14.120951	0.491445	41.888195	1950.378026	0.770472	0.0924
std	0.0	2.328915	2.960605	0.499954	13.992076	1189.294126	0.420552	0.2896
min	2019.0	0.000000	0.000000	0.000000	16.000000	256.000000	0.000000	0.0000
25%	2019.0	6.000000	12.000000	0.000000	30.000000	900.000000	1.000000	0.0000
50%	2019.0	7.000000	14.000000	0.000000	42.000000	1764.000000	1.000000	0.0000
75%	2019.0	10.000000	16.000000	1.000000	54.000000	2916.000000	1.000000	0.0000
max	2019.0	11.000000	22.000000	1.000000	89.000000	7921.000000	1.000000	1.0000

4.2.

Scatter plot ln(incwage) and education. Include a linear fit line. Be sure to label all axes and include an informative title.

In [86]:
```
plt.scatter(acs_data["educdc"], acs_data["incwage_log"], alpha = 0.1)
sn.regplot(x="educdc", y="incwage_log", data = acs_data)
plt.title("Wages tend to rise with greater education")
plt.xlabel("Education")
plt.ylabel("Income (log dollars)")
```

Out[86]:
```
Text(0, 0.5, 'Income (log dollars)')
```

4.3

Estimate a linear model of ln incwage and report your results

In [87]:
```
income_log_lm = smf.ols("incwage_log ~ educdc + female + AGE + age_squared + white + bl

print(income_log_lm.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            incwage_log   R-squared:                       0.331
Model:                            OLS   Adj. R-squared:                  0.330
Method:                 Least Squares   F-statistic:                     459.4
Date:                Mon, 01 Feb 2021   Prob (F-statistic):               0.00
Time:                        23:18:57   Log-Likelihood:                -13182.
No. Observations:                9293   AIC:                         2.639e+04
Df Residuals:                    9282   BIC:                         2.646e+04
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      4.8680      0.107     45.347      0.000       4.658       5.078
educdc         0.1122      0.004     30.819      0.000       0.105       0.119
female        -0.4552      0.021    -21.513      0.000      -0.497      -0.414
AGE            0.1918      0.005     37.368      0.000       0.182       0.202
age_squared   -0.0020   6.03e-05    -32.492      0.000      -0.002      -0.002
white         -0.0695      0.031     -2.258      0.024      -0.130      -0.009
black         -0.2771      0.045     -6.189      0.000      -0.365      -0.189
hispanic       0.0038      0.039      0.099      0.921      -0.073       0.080
married       -0.0467      0.090     -0.521      0.602      -0.223       0.129
NCHILD         0.0255      0.010      2.498      0.013       0.005       0.046
vet           -0.0077      0.049     -0.158      0.875      -0.103       0.088
==============================================================================
Omnibus:                     2632.721   Durbin-Watson:                   1.850
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10593.142
Skew:                          -1.354   Prob(JB):                         0.00
Kurtosis:                       7.475   Cond. No.                     2.39e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifi
```

```
ed.
[2] The condition number is large, 2.39e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

4.3.a.

What fraction of the variation in log wages does the model explain? Adjusted $R^2 = 0.330$

4.3.b.

Test the hypothesis that H0 : β1 = β2 = ... = β11 = 0 HA : βj 6= 0 f or some j with α = 0.10.

We can reject the null hypothesis at the alpha = 0.1 level. The F-statistic is 459.4, corresponding to a p value < 0.01. This means that it is very likely there is at least one predictor with a nonzero effect on log income.

4.3.c.

What is the return to an additional year of education? Is this statistically significant? Is it practically significant? Briefly explain.

An additional year of education results in an 11.2% increase in income. This effect is statistically significant at the p<0.05 level.

However, it is still not clear what this tells us, because much of the value of education is in the degrees attained. And there are cases in the data where a person might have gone through the full duration of a degree, but not achieved the degree. For instance, 4 years of college (with no degree) and a Bachelor's degree have the same value in educdc. This makes me wonder whether we're missing the impact of that degree on income, limiting the practical significance of our finding.

4.3.d.

At what age does the model predict an individual will achieve the highest wage?

To do this I maximize $log(income) = 15 - 0.1918Age + 0.0022Age^2$. Doing this, we receive a convex parabola with a minimum at 43.6. This seems to indicate that people have their lowest incomes around age 43; but the model predicts that the farther someone gets from this age, the higher their income will be. So, according to the model, income rises without boundary as a person grows older.

4.3.e.

Does the model predict that men or women will have higher wages, all else equal? Briefly explain why we might observe this pattern in the data.

The coefficient on the dummy variable "female" is -0.4552. This means our model predicts that women have 45.52% lower wages than men, all else equal. This reflects the "gender pay gap," in which women make less on average than men. There could be a number of causes, among them discrimination and the career/earnings interruptions women experience due to pregnancy and childcare.

4.3.f.

Interpret the coefficients on the white, black, and hispanic variables.

The coefficient on white indicates that being white has the effect of lowering income by 6.95%. Being black is associated with a 27.71% decrease in income, and being Hispanic is associated with a 0.38% increase in income. However, only the effects of white and black are significant at the p<0.05 level.

(g) Test the hypothesis that race has no effect on wages. Be sure to explicitly state the null and alternative hypotheses and show your calculations. H_0: β_white=β_black=β_hispanic=0 H_a: at least one β_j is non-zero

To test this hypothesis, I applied the formula:

$$F = \frac{RSS_0 - RSS/q}{RSS/(n-p-1)}$$

$$F = \frac{13741 - 9284/7}{9284/(9282)}$$

$$F = 636.46$$

Based on this very high F score, we can reject the null hypothesis. There is at least one race variable with an effect on income.

In [88]:
```
income_race_log_lm = smf.ols("incwage_log ~ white + black + hispanic", data = acs_data)
print(income_race_log_lm.summary())

RSS = income_log_lm.ssr
RSS_0 = income_race_log_lm.ssr
print("RSS = ", RSS, "\nRSS_0 = ", RSS_0)
print("F = ",((RSS_0-RSS)/7)/(RSS/(9282)))

print("COMPARISON\n",income_log_lm.compare_f_test(income_race_log_lm))
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            incwage_log   R-squared:                       0.010
Model:                            OLS   Adj. R-squared:                  0.010
Method:                 Least Squares   F-statistic:                     31.23
Date:                Mon, 01 Feb 2021   Prob (F-statistic):           4.43e-20
Time:                        23:19:02   Log-Likelihood:                -15003.
No. Observations:                9293   AIC:                         3.001e+04
Df Residuals:                    9289   BIC:                         3.004e+04
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
```

```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     10.4539      0.035    297.946      0.000      10.385      10.523
white         -0.0141      0.037     -0.378      0.705      -0.087       0.059
black         -0.3604      0.054     -6.635      0.000      -0.467      -0.254
hispanic      -0.2872      0.046     -6.199      0.000      -0.378      -0.196
==============================================================================
Omnibus:                     2179.326   Durbin-Watson:                   1.602
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5749.283
Skew:                          -1.261   Prob(JB):                         0.00
Kurtosis:                       5.913   Cond. No.                         6.95
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifi
ed.
RSS =  9284.30853124208
RSS_0 =  13740.670377274844
F =  636.4648253507474
COMPARISON
 (636.4648253507473, 0.0, 7.0)
```

4.

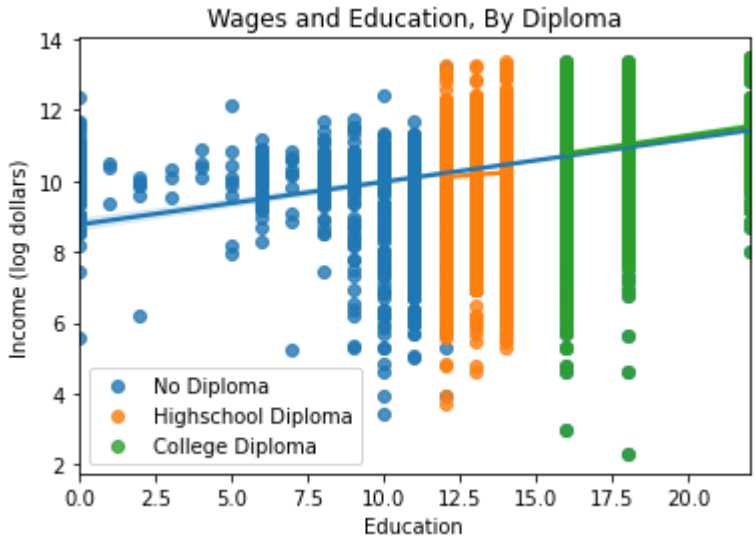
Graph ln(incwage) and education. Include three distinct linear fit lines specific to individuals with no high school diploma, a high school diploma, and a college degree. Be sure to label all axis and include an informative title.

NOTE: I notice that there is an outlier group in the no highschool group who nevertheless have many years of school. I believe these to be veterans eligible for certain types of PHD programs.

In [89]:
```
no_highschool = acs_data.query("hsdip == 0")
highschool = acs_data.query("hsdip == 1")
college = acs_data.query("coldip == 1")

sn.regplot(x="educdc", y="incwage_log", data = no_highschool, label = "No Diploma")
sn.regplot(x="educdc", y="incwage_log", data = highschool, label = "Highschool Diploma"
sn.regplot(x="educdc", y="incwage_log", data = college, label = "College Diploma")
plt.title("Wages and Education, By Diploma")
plt.xlabel("Education")
plt.ylabel("Income (log dollars)")
plt.legend()
```

Out[89]:
```
<matplotlib.legend.Legend at 0x1fb0803e8b0>
```



6.

Estimate the model you proposed in the previous question and report your results.

In [90]:
```
income_log_lm_expanded = smf.ols("incwage_log ~ educdc + female + AGE + age_squared + w
print(income_log_lm_expanded.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            incwage_log   R-squared:                       0.350
Model:                            OLS   Adj. R-squared:                  0.349
Method:                 Least Squares   F-statistic:                     357.3
Date:                Mon, 01 Feb 2021   Prob (F-statistic):               0.00
Time:                        23:19:08   Log-Likelihood:                -13046.
No. Observations:                9293   AIC:                         2.612e+04
Df Residuals:                    9278   BIC:                         2.623e+04
Df Model:                          14
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept        6.1770      0.136     45.387      0.000       5.910       6.444
educdc          -0.0205      0.009     -2.197      0.028      -0.039      -0.002
female          -0.4650      0.021    -22.238      0.000      -0.506      -0.424
AGE              0.1775      0.005     34.527      0.000       0.167       0.188
age_squared     -0.0018   6.02e-05    -29.964      0.000      -0.002      -0.002
white           -0.0393      0.030     -1.291      0.197      -0.099       0.020
black           -0.2098      0.044     -4.723      0.000      -0.297      -0.123
hispanic        -0.0002      0.039     -0.004      0.997      -0.076       0.076
married         -0.0006      0.088     -0.007      0.995      -0.174       0.173
NCHILD           0.0280      0.010      2.772      0.006       0.008       0.048
vet              0.0057      0.048      0.118      0.906      -0.089       0.100
hsdip           -0.7331      0.213     -3.434      0.001      -1.152      -0.315
coldip          -0.6302      0.211     -2.985      0.003      -1.044      -0.216
educdc:hsdip     0.1047      0.018      5.947      0.000       0.070       0.139
educdc:coldip    0.1154      0.015      7.871      0.000       0.087       0.144
==============================================================================
Omnibus:                     2799.529   Durbin-Watson:                   1.899
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            11886.299
Skew:                          -1.426   Prob(JB):                         0.00
Kurtosis:                       7.750   Cond. No.                     5.33e+04
==============================================================================
```

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specifi
ed.
[2] The condition number is large, 5.33e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Having a highschool diploma is associated with a 73.31% decrease in income. However, it also has an interaction effect with duration of education: every additional year of education increases income by 10.5%. Meanwhile, having a college diploma is associated with a 63% decrease in income compared to having no degree. But it has an even stronger interaction with education; every additional year of education increases a college diploma holder's predicted income by 11.5%. All of these effects are significant at the $p<0.05$ level.

6.a.

Predict the wages of an 22 year old, female individual (who is neither white, black, nor Hispanic, is not married, has no children, and is not a veteran) with a high school diploma and an all else equal individual with a college diploma. Assume that it takes someone 12 years to graduate high school and 16 years to graduate college.

In [92]:
```
person_no_college_dict = {"female":[1],
          "AGE":[22],
          "age_squared":[484],
          "white":[0],
          "black":[0],
          "hispanic":[0],
          "married":[0],
          "NCHILD":[0],
          "vet":[0],
          "hsdip":[1],
          "coldip":[0],
          "educdc":[12]}
person_no_college = pd.DataFrame(data=person_no_college_dict)

person_college_dict = {"female":[1],
          "AGE":[22],
          "age_squared":[484],
          "white":[0],
          "black":[0],
          "hispanic":[0],
          "married":[0],
          "NCHILD":[0],
          "vet":[0],
          "hsdip":[0],
          "coldip":[1],
          "educdc":[16]}
person_college = pd.DataFrame(data=person_college_dict)

prediction_no_college = income_log_lm_expanded.get_prediction(person_no_college)
print("No College: $", math.exp(prediction_no_college.summary_frame(alpha=0.05)["mean"]

prediction_college = income_log_lm_expanded.get_prediction(person_college)
print("College: $", math.exp(prediction_college.summary_frame(alpha=0.05)["mean"]))
```

```
No College: $ 8276.907068458813
College: $ 15239.420041477411
```

6.b.

The President wants to know, given your results, do individuals with college degrees have higher predicted wages than those without? By how much? Briefly explain.

Yes, individuals with college degrees have higher predicted wages than those without. While the direct effect of the diploma is negative, the interaction effect with years of education is more important. For instance, when an individual graduates they receive a diploma and they also have 12 years of education. So although we predict that having the diploma in isolation lowers her income 63%, the interaction effect adds about 138% to her predicted income. So the net effect greatly increases her income.

6.c.

The President asked you to look into this question because she is considering legislation that will expand access to college education (for instance, by increasing student loan subsidies). She will only support the legislation if there are cost offsets (if college education increases wages and therefore, future income tax revenues that help reduce the net cost of the subsidy). Given that criteria, how would you advise the President?

Based on the reasoning presented in 6.b., there is a strong case for this legislation. The cost of expanding college is likely to be offset by the very signifant increases to future earnings.

However, if I were being more honest (for instance, working behind the scenes with the president's team of policy advisors), I would advise some additional caution. Even though we have disentangled the effects of years of education and receipt of diploma, we don’t know many of the factors that determine whether someone is likely to receive a diploma. For instance, it could be that individuals who attain college diplomas are also more likely to have high income parents. If that’s the case, the observed effect may be due to the positive effects of high parental income rather than from attaining a diploma.

7.

There are many ways that this model could be improved. How would you do things differently if you were asked to predict the returns to education given the data available on IPUMS?

I would include additional variables in the regression, being careful to articulate clear theoretical justifications for each and employing a variable selection algorithm to help keep my list of variables to a reasonable size. Some IPUMS variables that seem promising include: URBAN Urban/rural status CITIZEN Citizenship status YRSUSA1 Years in the United States DEGFIELD Field of degree

I would also recommend we make use of IPUM's weights to ensure that our sample is representative.