Wasil Engel
12231558
Lab Mini Project 1

I partnered up with Ricardo Saucedo (12245077).

To see more than the screenshots of the plots I created for the coding part of the assignment, please see my Jupyter notebook on Github here: https://github.com/Wasil-UChi/Machine-Learning/tree/master/PUMS

#1 Overview

- Work for the President's Council of Economic Advisors (CEA) – executive branch, setting economic policy
- Here, data from the American Community Survey (ACS) Public Use Microdata Sample (PUMS)
- Goal: predict returns to education -> set education agenda with a specific focus on the expansion of access to higher education

#2 Obtaining the Data

#3 Preparing the Data

```
# 1

path_1 = '/Users/wasilengel/Desktop/School/Harris/Machine Learning/PUMS/usa_00001.csv'

path_2 = '/Users/wasilengel/Desktop/School/Harris/Machine
Learning/PUMS/PPHA_30545_MP01-Crosswalk.csv'

df = pd.read_csv(path_1)

df.head()

crosswalk = pd.read_csv(path_2)

crosswalk



# 2a
```

```
# vector = np.array(crosswalk["educdc"])
# vector

crosswalk_dict = dict(zip(crosswalk.educd, crosswalk.educdc))
crosswalk_dict

df["EDUCD"].unique()

df["educdc"] = df["EDUCD"]
df["educdc"].replace(crosswalk_dict, inplace = True)
df.head()



# 2b

# i

df["hsdip"] = df["educdc"]==12.0
df["hsdip"] = df["hsdip"].astype(int)
df.head()

# ii

df["coldip"] = df["educdc"]==16.0
df["coldip"] = df["coldip"].astype(int)
df.head()

# iii

df["white"] = df["RACE"]==1
df["white"] = df["white"].astype(int)
df.head()

# iv

df["black"] = df["RACE"]==2
df["black"] = df["black"].astype(int)
df.head()

# v

# df["MARST"].unique() shows there are no nines!
```

```
df["hispanic"] = df["HISPAN"]!=0
df["hispanic"] = df["hispanic"].astype(int)
df.head()

# vi

df["married"] = (df["MARST"]==1) | (df["MARST"]==2)
df["married"] = df["married"].astype(int)
df.head()

# vii

df["female"] = df["SEX"]==2
df["female"] = df["female"].astype(int)
df.head()

# viii

df["vet"] = df["VETSTAT"]==2
df["vet"] = df["vet"].astype(int)
df.head()



# 2c

df["educdc X hsdip"] = df["educdc"] * df["hsdip"]
df.head()
# Makes sense intuitively: howing us that person got 12 years of education

df["educdc X coldip"] = df["educdc"] * df["coldip"]
df.head()
# Makes sense intuitively: showing us that person got 16 years of education



# 2d

# i

df["AGEsq"] = df["AGE"] * df["AGE"] # or, df["AGE"]^2 (same results)
df.head()

# ii
```

```
df["lnincwage"] = np.log(df["INCWAGE"])
df = df[df["lnincwage"] > 1]
df.head()
# df["lnincwage"].unique() # see that -inf is now being dropped
```

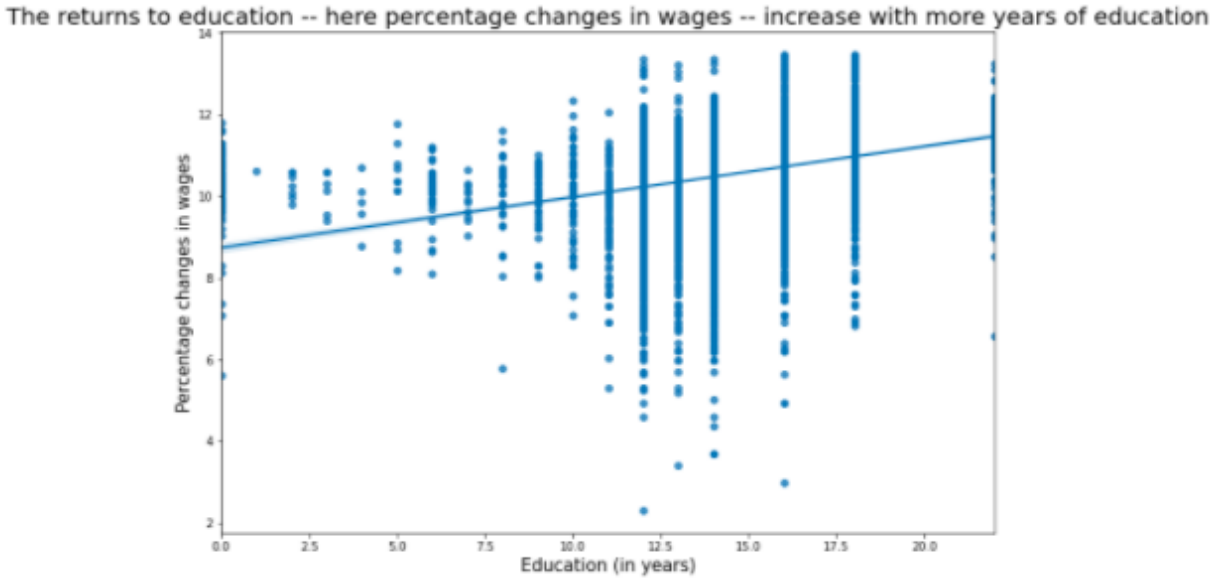
#4 Performing Data Analysis & Answering Questions

```
# 1

df.columns

YEAR_sum = df["YEAR"].describe()
YEAR_sum
INCWAGE_sum = df["INCWAGE"].describe()
INCWAGE_sum
lnincwage_sum = df["lnincwage"].describe()
lnincwage_sum
educdc_sum = df["educdc"].describe()
educdc_sum
female_sum = df["female"].describe()
female_sum
AGE_sum = df["AGE"].describe()
AGE_sum
AGEsq_sum = df["AGEsq"].describe()
AGEsq_sum
white_sum = df["white"].describe()
white_sum
black_sum = df["black"].describe()
black_sum
hispanic_sum = df["hispanic"].describe()
hispanic_sum
married_sum = df["married"].describe()
married_sum
NCHILD_sum = df["NCHILD"].describe()
NCHILD_sum
vet_sum = df["vet"].describe()
vet_sum
hsdip_sum = df["hsdip"].describe()
hsdip_sum
coldip_sum = df["coldip"].describe()
coldip_sum
educdcXhsdip_sum = df["educdc X hsdip"].describe()
educdcXhsdip_sum
```

```
educdcXcoldip_sum = df["educdc X coldip"].describe()
educdcXcoldip_sum
```

2

```
plt.figure(figsize=(12,8))
sns.regplot(x="educdc", y='lnincwage', data=df)
plt.xlabel("Education (in years)", fontsize=15)
plt.ylabel("Percentage changes in wages", fontsize=15)
plt.title("The returns to education -- here percentage changes in wages -- increase with more years of education", fontsize=20)
plt.show()
```



3

```
df.columns
```

```
# import statsmodels.formula.api as smf
result = smf.ols('lnincwage ~ educdc + female + AGE + AGEsq + white + black + hispanic + married + NCHILD + vet', data = df).fit()
print(result.summary())
```

a

The fraction of the variation explained is given by R-squared value (the R-squared adjusted value for the additional predictors is very similar so I'll go with the more prevalent one used in academia, the regular R-squared). The value is 0.306 hence the model explains 30.6 per cent of the variation in our depenedent variable, log wages.

b

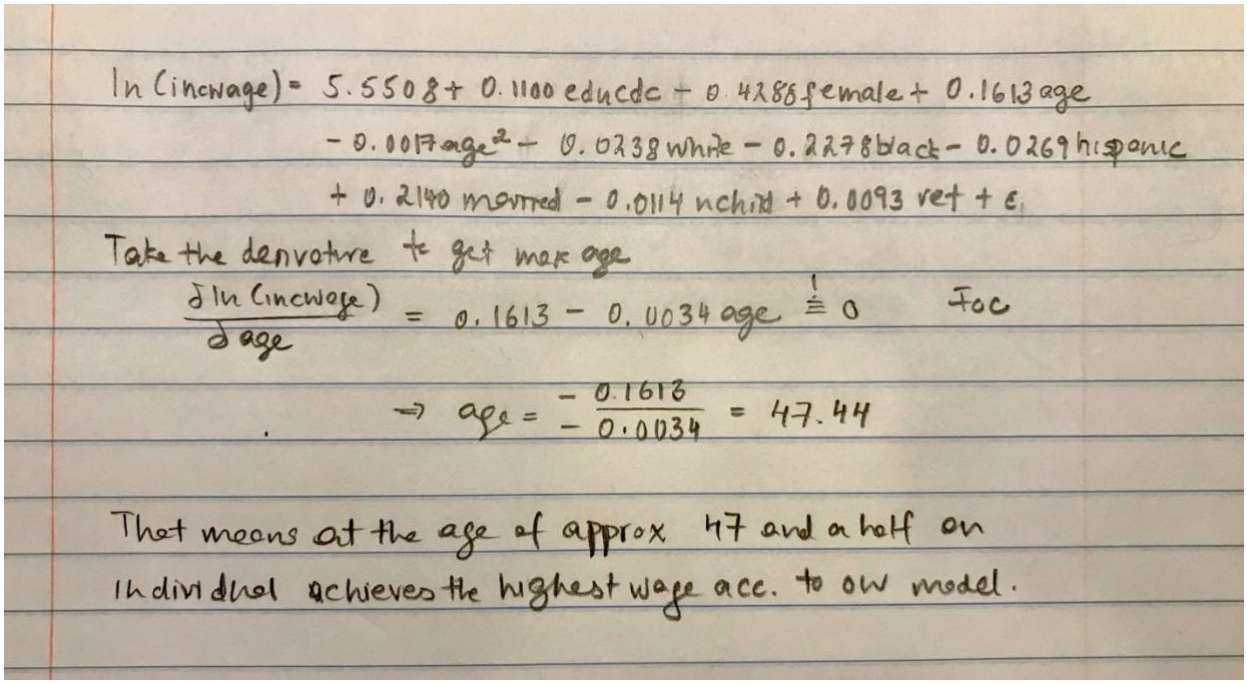
The null is very pessimistic ("none of our Xs combined have predictive value for wages"): the one we look for is the F-test! The F-statistic is at 380.3, that's a very large value. Consequently, the p-value is very low as we can see: 0.00. That said, we can definitely reject the null at a 10% significance level, we could even reject it at a more conservative level too, e.g. 5% or 1%. Hence, we can conclude that all of our coefficients are JOINTLY significant, hence, this model with the Xs chosen does a good job at predicting the wages.
This is not be confused with the respective p-values from a bivariate standpoint that is shown in each row pertaining to the respective independent variables, the Xs. As we can see there, from the regression table, all of our independent values are statistically significant too EXCEPT for: white, hispanic, NCHILD, and vet. Yet, this is different from the F-statistic which as explained in the line above takes the significance of the model as a whole, with all included variables being tested JOINTLY.

c

Because we are using log wages, an additional year of education yields an 11 per cent increase in log wages, ceteris paribus. As we can see from the corresponding p-value, this is highly significant, even at a conservative threshold.
Talking about the practical significane, hence the magnitude of the coefficient, the practical signifance will depend upon the actual income of a person in real wages. For example, if a person earns $10, an addtional dollar earned (rounded) means more to him/ her than a wage increase from 100,000 to 110,000 dollars for someone who already is rich and can afford practically anything. From an economic point of view, however, an increase of 10%, in the example of the rich person I gave that is $10,000, is a lot of money and hence meaningful.
That said, given the magnitude, 11%, I would definitely argue that the returns of an additional year of education in terms of wage increases are not just statistically but also practically/ economically meaningful: they do have quite a large significance, both in statistical and practical terms.

d

Please see hand-written snap in WORD document. The answer is that my model predicts at an age of approx. 47 and a half an individual achieves the highest wage.



e

Given that our female variable is a dummy and the highly statistically coefficient at -0.4288, ceteris paribus, we can conclude that being a women affects log wages negatively whereas being a man (code female = 0), doesn't, ceteris paribus. Hence being a man doesn't affect the wages negatively while being a women, it does at a rate of approx. -43% (which also is the percentage difference between the two holding all other Xs constant).

f

The coefficient for white is at -0.0238, ceteris paribus, suggesting that being white affects wages negatively at approx. -2.4 per cent. However, it's not statistically significant, we cannot reject the null for white. Unlike for black, which is highly significant. The coefficient is at -0.2278, holding all else equal, and thus implies being black negatively affects the log wages at a rate of approx. -22.8 per cent (22.8 per cent less pay on average when being black!). Similarly to whites, the coefficient suggests that being hispanic negatively impacts wages at -0.0269 meaning on average 2.7 per cent less wages, however, the hispanic predictor is not statistically significant either holding all else equal.

g

I do not assume that being hispanic is a race. Hence, I will test for black and white only. Note: F-Test!
That said, the null is that beta5 = beta6 = 0 meaning race has no effect on wages hence coding white or black as 1 (it's a dummy) will result in a zero percent change in log wages under the null. As common in the social sciences, I'll use an alpha of 5 percent (=0.05).

The alternative (I take a two-sided approach) states that the effect of race on wages is non zero hence affecting the log wages either positive or negative which makes sense here because we test for both white and black with effects going in diffeent directions (assuming wages go up for whites and down for blacks due to racial injustice).
That said, I'd assume the effects would cancel each other out (that's for the model as a whole! -> taken jointly: F-statistic!) and our results would be inconclusive (for the model as a whole) hence not statistically significant. But let's see ...

```
result = smf.ols('lnincwage ~ white + black', data = df).fit()
print(result.summary())
```

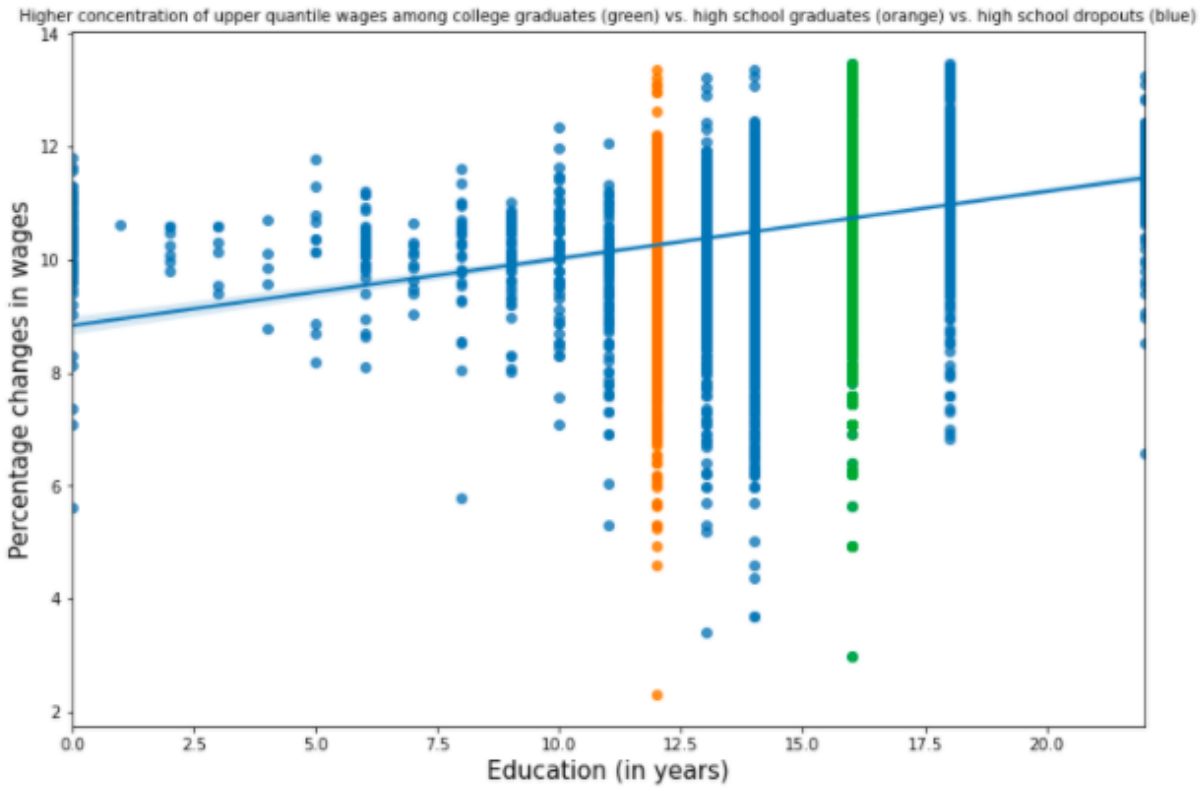
First note how low the R-squared value is: this model does only explain 1 per cent of the variation of logwages! That means there's so much more to explaining the changes in wages than just race. In any case:
My results are consistent with my previous findings and also with my assumptions.
Looking at my findings from a joint perspective, meaning race as stated in my null, I find that the F-statistic is rather low, 44.51, and the corresponding p-level at 5.8 per cent would NOT pass the common 5 percent threshold. Hence, I would fail to reject the null at a 5 per cent significance level and conclude that the coefficients are not JOINTLY significant.

But are they individually? Note: this bit goes beyond my null but since it further strengthens my findings, I shall include these following observations too.
Breaking my findings down further and looking at each X individually now, I can see that:
Being white increases logwages by approx. 5.6 per cent while being black leads to a wage decrease of approx. 33.7%. However, given that only one of the two is statsitcially significant (black is at a low threshold/ a significance level of 1%, being white isn't significant not even at a 10% significance level),
this further supports my previous findings where I tested for both at the same time: race alone does not affect wages and as the R-squared suggests, there's more to it than just race. We would need to include more regressors because as it stands now, our model suffers from Omitted Variable Bias.

4

```
#import numpy as np
#from numpy.polynomial.polynomial import polyfit
#import matplotlib.pyplot as plt

x_1 = df[df["hsdip"]==0]
x_2 = df[df["hsdip"]==1]
x_3 = df[df["coldip"]==1]
x_1 = x_1[["lnincwage", "educdc", "hsdip"]]
x_2 = x_2[["lnincwage", "educdc", "hsdip"]]
```

```
x_3 = x_3[["lnincwage", "educdc", "coldip"]]
x_1

# The orange line is at 12 years so hsdip = 1 and the green one is at 16 years so coldip = 1
# It makes sense for these two fitted lines to be vertical because at 12 or 16 years there is no percentage change because we don't move along the x-axis (no 1-unit increase or decrease that could result in a wage change).

plt.figure(figsize=(12,8))
sns.regplot(x="educdc", y='lnincwage', data=x_1)
sns.regplot(x="educdc", y='lnincwage', data=x_2)
sns.regplot(x="educdc", y='lnincwage', data=x_3)
plt.xlabel("Education (in years)", fontsize=15)
plt.ylabel("Percentage changes in wages", fontsize=15)
plt.title("The returns to education increase with more years of education: higher concentration of upper quantile wages among college graduates (green) vs. high school graduates (orange) vs. high school dropouts (blue)", fontsize=20)
plt.show()

# Or:
# x_1
#plt.scatter(x_1["educdc"], x_1["lnincwage"])
#m, b = np.polyfit(x_1["educdc"], x_1["lnincwage"], 1)
#plt.plot(x_1["educdc"], m*x_1["educdc"] + b)

# x_2
#plt.plot(x_2["educdc"], x_2["lnincwage"], 'o')
#m, b = np.polyfit(x_2["educdc"], x_2["lnincwage"], 1)
#plt.plot(x_2["educdc"], m*x_2["educdc"] + b)

# x_3
#plt.plot(x_3["educdc"], x_3["lnincwage"], 'o')
#m, b = np.polyfit(x_3["educdc"], x_3["lnincwage"], 1)
#plt.plot(x_3["educdc"], m*x_3["educdc"] + b)
```



5

```
# import statsmodels.formula.api as smf
result = smf.ols('lnincwage ~ coldip + hsdip + educdc + female + AGE + AGEsq + white + black + hispanic + married + NCHILD + vet', data = df).fit()
print(result.summary())
```

This model explains slightly more of the variation in y now that I included the two dummies up front: R^2 rises slightly to almost 31% and the findings make sense with regards to what I uncovered previously and assumed:
Both dummies are statistically significant at a 5-% significance level. Having a college degree (code as coldip = 1) will increase wages by 14 per cent (wow!), ceteris parivbus. Having no high school diploma decrreass wages by approx. 6 per cent, which is also the difference in expected wages when having graduated from high school.
That said, having a college degree is not only a stronger predictor for wages, much more so than having a high school diploma judging by the higher p-value, the magnitude of the coefficient also suggests that its impact on wages is also much more meaningful: an increase by 14 per cent!

6

To see more than the screenshots of the plots I created for the coding part of the assignment, please see my Jupyter notebook on Github here: *https://github.com/Wasil-UChi/Machine-Learning/tree/master/PUMS*

```
# a

individual_1 = {'AGE': [22], "female": [1], "white": [0], "black": [0], "hispanic": [0], "married":
[0], "NCHILD": [0], "vet": [0], "hsdip": [1], "coldip": [0], "educdc": [12], "AGEsq": [484]}
predict_1 = pd.DataFrame(data=individual_1)
print(predict_1)
prediction_1 = result.get_prediction(predict_1)
prediction_1.summary_frame(alpha=0.05)

# What is the predicted wage in absolute terms associated with a log wage of 9.124729?
# from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(df["lnincwage"].to_frame(), df["INCWAGE"])
model.predict(pd.Series([9.124729]).to_frame())

individual_2 = {'AGE': [22], "female": [1], "white": [0], "black": [0], "hispanic": [0], "married":
[0], "NCHILD": [0], "vet": [0], "hsdip": [1], "coldip": [1], "educdc": [16], "AGEsq": [484]}
predict_2 = pd.DataFrame(data=individual_2)
print(predict_2)
prediction_2 = result.get_prediction(predict_2)
prediction_2.summary_frame(alpha=0.05)

# What is the predicted wage in absolute terms associated with a log wage of 9.659923?
# from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(df["lnincwage"].to_frame(), df["INCWAGE"])
model.predict(pd.Series([9.659923]).to_frame())

# Compare the findings with our summary stats from #1:
lnincwage_sum

INCWAGE_sum

# b

# To begin with, see answer to # 5 where I wrote:
# "Both dummies are statistically significant at a 5-% significance level. Having a college degree
(code as coldip = 1) will increase wages by 14 per cent (wow!), ceteris parivbus. Having no high
school diploma decreases wages by approx. 6 per cent, which is also the difference in expected
wages when having graduated from high school.
```

That said, having a college degree is not only a stronger predictor for wages, much more so than having a high school diploma judging by the higher p-value, the magnitude of the coefficient also suggests that its impact on wages is also much more meaningful: an increase by 14 per cent!"

In addition to that, based on my findings in 6a, I can say that:
Yes, individuals with college degrees do have higher predicted wages than those without.
In our specific example (for general interpretation, see above), the first individual with a high school diploma but no college degree earns on average $555.4 less than what is expected (note: I converted to absolute absolute terms for easier interpretation for the President).
That changes for the very same individual with equal characterics except for that she now has a college degree. She earns now on average $22686 more than expected.
Taken together, the difference between the two is at approx. $23423 (22868 + 555) in absolute terms.
In relative terms, the wage increases by approx. 54 per cent for the specific individual we looked at, that is: 9.66-9.12 (note both are random variables in themselves).

So, yes, Mr. President, both with regards to both the general interpretation in #5 as well as #6a, both in absolute and relative/ comparative terms, individuals holding a college degree do earn more on average than individuals without, holding all other factors constant.

c

Given what we just learned, and in addition to my elaborations in 6b and 5, I would advise the President to implement the policy because the returns to education are notable as elaborated above in 6b.
General: +14 per cent -> that is the increase holding a college degree vs. holding none, holding all others factors constant
Individual: +54 per cent -> that is the increase specific to a young female
That in addition to a progressive tax policy in which people who earn more, pay more, is likely to offset the cost of the subsidies and on top of that disproportionally empowers females, which is desirable and needed.
Why needed? Because that individual still, despite the college degree, earns less than average: she's at $22686 and the median is $41500 (I don't consider the mean because income data is highly skewed). Yet, it's a step in the right direction, so yes, Mr. President, good idea, education is the grand equalizer, however depending on the purpose of the policy, say promoting gender equality, it can be further enhanced by disproportionaly promoting young females (or people of color), e.g. offering them tax deductions in the future if they go to college. However, if the actual purpose is to secure a budget increase in the future, then this would be counterintuitive.
That said, good idea, but I'll need to make sure to get more information from the President in order to come to a definite conclusion.

7

There are a few ways to do things differently, e.g.
- Include other or more regressors, e.g. first generation immigrants, state/ geography, handicaps, etc.
- Use a non-linear model (that we are yet to learn about in class for quantitative outcomes)
- Instead having income as the dependent variable, take e.g. likelihood of debt as a measure for financial security later in life and run a Linear Probability Model (LPM) OR a Quadratic Discriminat Analysis which allows for many Xs and is non-linear