In [5]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import cross_val_score
import sklearn.metrics
from sklearn.preprocessing import PolynomialFeatures
from sklearn.model_selection import train_test_split
import statsmodels.api as sm
sns.set()  #this allows you to use seaborn themes with matplotlib functions
rand_state= 1000
dtafile = "D:/CGO/Migration Research with Prof. Max/GSS/gss2021_born=2.dta"
df = pd.read_stata(dtafile)
df.tail()
```

```
C:\Users\aflat\anaconda3\lib\site-packages\pandas\io\stata.py:1417: UnicodeWarning:
One or more strings in the dta file could not be decoded using utf-8, and
so the fallback encoding of latin-1 is being used.  This can happen when a file
has been incorrectly encoded by Stata or some other software. You should verify
the string values returned are correct.
  warnings.warn(msg, UnicodeWarning)
```

Out[5]:

	year	id	wrkslf	wrkgovt	marital	martype	divorce	widowed	age	agekdbrn	...	relitennv	biblenv	postlifenv	kidssolnv	uscitznnv	fucitznnv	fepolnv	scibnftsnv	abanyg	f
439	2021	4390	NaN	NaN	1.0	NaN	2.0	2.0	NaN	33.0	...	NaN	3.0	2.0	3.0	1.0	NaN	2.0	1.0	NaN	
440	2021	4427	NaN	NaN	1.0	NaN	2.0	2.0	29.0	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
441	2021	4430	NaN	NaN	5.0	NaN	NaN	NaN	25.0	NaN	...	1.0	2.0	1.0	1.0	2.0	1.0	NaN	2.0	1.0	
442	2021	4459	NaN	NaN	1.0	NaN	2.0	2.0	81.0	22.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
443	2021	4469	NaN	NaN	1.0	NaN	2.0	2.0	NaN	34.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

5 rows × 565 columns

In [6]:

```
import pandas as pd
from sklearn import linear_model
import statsmodels.api as sm
from scipy import stats

#RACEACS1','RACEACS2', 'RACEACS3', 'RACEACS3

X = df[['born', 'raceacs1', 'raceacs2', 'raceacs3', 'raceacs4', 'raceacs5', 'raceacs6', 'raceacs7', 'raceacs8', 'raceacs9', 'raceacs10', 'raceacs11
Y = df['immlimit']
Z=df[[]]
X=X.fillna(0)
Y=Y.fillna(0)
print(X.shape)
print(Y.shape)
# with sklearn
#regr = linear_model.LinearRegression()
#regr.fit(X, Y)

# with statsmodels
X = sm.add_constant(X) # adding a constant

model = sm.OLS(Y, X).fit()
predictions = model.predict(X)

print_model = model.summary()
print(print_model)
# plt.plot(X,Y,'o')
# plt.plot(X, model.fittedvalues)
```

```
(444, 19)
(444,)
                            OLS Regression Results
==============================================================================
Dep. Variable:               immlimit   R-squared:                       0.566
Model:                            OLS   Adj. R-squared:                  0.551
Method:                 Least Squares   F-statistic:                     37.22
Date:                Tue, 02 Aug 2022   Prob (F-statistic):           5.53e-68
Time:                        16:53:37   Log-Likelihood:                -724.08
No. Observations:                 444   AIC:                             1480.
Df Residuals:                     428   BIC:                             1546.
Df Model:                          15
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
born           0.1105      0.166      0.664      0.507      -0.216       0.437
raceacs1       0.3755      0.323      1.162      0.246      -0.260       1.011
raceacs2       0.2048      0.355      0.577      0.564      -0.492       0.902
raceacs3      -0.6850      0.896     -0.764      0.445      -2.447       1.077
raceacs4       0.3307      0.376      0.881      0.379      -0.407       1.069
raceacs5       0.5987      0.397      1.509      0.132      -0.181       1.378
raceacs6       0.0243      0.554      0.044      0.965      -1.064       1.112
raceacs7       0.6150      0.656      0.938      0.349      -0.674       1.904
raceacs8       0.6628      0.443      1.496      0.135      -0.208       1.534
raceacs9       0.1887      0.949      0.199      0.842      -1.676       2.054
raceacs10      0.2653      0.354      0.750      0.453      -0.430       0.960
raceacs11  -3.699e-16   5.69e-16     -0.650      0.516   -1.49e-15    7.49e-16
raceacs12   6.185e-17   4.75e-16      0.130      0.896   -8.72e-16    9.96e-16
raceacs13  -1.878e-16   4.29e-16     -0.437      0.662   -1.03e-15    6.56e-16
raceacs14      0.4548      0.949      0.479      0.632      -1.411       2.320
raceacs15      0.3011      0.493      0.611      0.542      -0.668       1.270
raceacs16      0.1933      0.345      0.559      0.576      -0.486       0.872
trmedia        0.5293      0.023     23.054      0.000       0.484       0.574
realinc      9.34e-08   1.39e-06      0.067      0.947   -2.65e-06    2.83e-06
==============================================================================
Omnibus:                       79.025   Durbin-Watson:                   2.049
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              163.408
Skew:                           0.959   Prob(JB):                     3.28e-36
Kurtosis:                       5.271   Cond. No.                     7.62e+22
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 2.78e-34. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Dependent variable

immlimit Variable: IMMLIMIT Type: Numeric Label: America should limit immigration in order to protect our national way of life. STRONGLY AGREE [1] AGREE [2] NEITHER AGREE NOR DISAGREE [3] DISAGREE [4] STRONGLY DISAGREE [5]

Inependent Variables

born - Were you born in this country? RACEACS1 1 What is your race? White
RACEACS2 1 Black or African American
RACEACS3 1 American Indian or Alaska Native
RACEACS4 1 Asian Indian
RACEACS5 1 Chinese
RACEACS6 1 Filipino
RACEACS7 1 Japanese
RACEACS8 1 Korean
RACEACS9 1 Vietnamese
RACEACS10 1 Other Asian
RACEACS11 1 Native Hawaiian
RACEACS12 1 Guamanian or Chamorro
RACEACS13 1 Samoan
RACEACS14 1 Other Pacific Islander
RACEACS15 1 Some other race
trmedia - Variable: TRMEDIA Type: Numeric
Label:(On a scale of 0 to 10,how much do you personally trust each of the following institutions? 0 means you do not trust an institution at all, and 10 means you trust it completely.) The news media
realinc - Variable: REALINC Type: Numeric Label: Family income in 1972-2006 surveys in constant dollars (base=1986) Ranges from under $1,000 to$74,999 in total 18 categories