Set up

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import statsmodels.formula.api as smf
from __future__ import division
from statsmodels.compat import lzip
import scipy as sp
```

```
!pip install linearmodels
from linearmodels.iv import IV2SLS
```

```
Collecting linearmodels
  Downloading linearmodels-5.2-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (1.8 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 1.8/1.8 MB 8.2 MB/s eta 0:00:00
Requirement already satisfied: numpy>=1.19.0 in /usr/local/lib/python3.10/dist-packages (from linearmodels) (1.23.5)
Requirement already satisfied: pandas>=1.1.0 in /usr/local/lib/python3.10/dist-packages (from linearmodels) (1.5.3)
Requirement already satisfied: scipy>=1.5.0 in /usr/local/lib/python3.10/dist-packages (from linearmodels) (1.11.2)
Requirement already satisfied: statsmodels>=0.12.0 in /usr/local/lib/python3.10/dist-packages (from linearmodels) (0.14.0
Collecting mypy-extensions>=0.4 (from linearmodels)
  Downloading mypy_extensions-1.0.0-py3-none-any.whl (4.7 kB)
Requirement already satisfied: Cython>=0.29.34 in /usr/local/lib/python3.10/dist-packages (from linearmodels) (3.0.2)
Collecting pyhdfe>=0.1 (from linearmodels)
  Downloading pyhdfe-0.2.0-py3-none-any.whl (19 kB)
Collecting formulaic>=0.6.1 (from linearmodels)
  Downloading formulaic-0.6.4-py3-none-any.whl (88 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 88.9/88.9 kB 8.1 MB/s eta 0:00:00
Collecting setuptools-scm[toml]<8.0.0,>=7.0.0 (from linearmodels)
  Downloading setuptools_scm-7.1.0-py3-none-any.whl (43 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 43.8/43.8 kB 3.7 MB/s eta 0:00:00
Collecting astor>=0.8 (from formulaic>=0.6.1->linearmodels)
  Downloading astor-0.8.1-py2.py3-none-any.whl (27 kB)
Collecting interface-meta>=1.2.0 (from formulaic>=0.6.1->linearmodels)
  Downloading interface_meta-1.3.0-py3-none-any.whl (14 kB)
Requirement already satisfied: typing-extensions>=4.2.0 in /usr/local/lib/python3.10/dist-packages (from formulaic>=0.6.1
Requirement already satisfied: wrapt>=1.0 in /usr/local/lib/python3.10/dist-packages (from formulaic>=0.6.1->linearmodels
Requirement already satisfied: python-dateutil>=2.8.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.1.0->lin
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-packages (from pandas>=1.1.0->linearmodels)
Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from setuptools-scm[toml]<8.0.
Requirement already satisfied: setuptools in /usr/local/lib/python3.10/dist-packages (from setuptools-scm[toml]<8.0.0,>=7
Requirement already satisfied: tomli>=1.0.0 in /usr/local/lib/python3.10/dist-packages (from setuptools-scm[toml]<8.0.0,>
Requirement already satisfied: patsy>=0.5.2 in /usr/local/lib/python3.10/dist-packages (from statsmodels>=0.12.0->linearm
Requirement already satisfied: six in /usr/local/lib/python3.10/dist-packages (from patsy>=0.5.2->statsmodels>=0.12.0->li
Installing collected packages: setuptools-scm, mypy-extensions, interface-meta, astor, pyhdfe, formulaic, linearmodels
Successfully installed astor-0.8.1 formulaic-0.6.4 interface-meta-1.3.0 linearmodels-5.2 mypy-extensions-1.0.0 pyhdfe-0.2
```

```
#create dataframe of variables of interest
d = pd.read_csv(('GSS_Cum.csv'),
usecols= ['workblks', 'health', 'realinc', 'year', 'cohort', 'race', 'racwork'])
```

```
#recode to make variables more interpretable
d['workblks'] = 7 - d.workblks
d['health'] = 4 - d.health
d['inc10k'] = d.realinc/10000
#remove all non-values
data= d.dropna()
```

Variables

I am interested in exploring whether being biased or racist impacts overall health. The logic being that if a person has a severe bias against a group of people, that will lead to higher cortisol levels whenever they are exposed to somone who is a part of that group, due to an increase in fear or frustration that they associate with that group. Having frequently high cortisol is deleterious to health. I am curious to see if this

hypothesized increase in cortisol would produce a causal relationship between a racism indicator and perceived health. If my assumption holds true, my independent variable of perceived level of how hardworking black people are will be predictive of worse overall health.

Independent Variable

My independent variable for this exercise is 'Workblks'. This variable has respondents ranking how hard working they believe black people to be from hard working (1) to lazy (7). I am using this a proxy for bias against black people which may indicate a level of racism that may lead to increase stress/cortisol levels as noted above. I recoded this variable to be increasing in perceived hard working-nes to be more intuitive to interpret, so higher numbers indicate more positive views.

Dependent Variable

My dependent variable for this excercise is 'health' which asks respondents to rank from execellent to poor what would say their own health is. I again recoded this variable to go from poor to excellent so that higher numbers indicate better health, making the variable more easily interpretable.

Controls

For some basic controls I included the year of the survey response, the cohort of the respondent (how old they are), and their race (white- 1, black - 2, or other -3). I also included real income which is the income in real dollars which I recoded to be in icrements of 10 thousand dollars. The assumptions are that richer individuals will have higher levels of health and that race will also impact how the person views black people. Age also would presumably impact health levels and also older people may be more likely to be more prejudiced.

OLS

```
# Fit the linear regression model
lm1 = smf.ols('health ~ workblks + inc10k + year + cohort + C(race)', data = data).fit()

# Print the model summary
print(lm1.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 health   R-squared:                       0.036
Model:                            OLS   Adj. R-squared:                  0.035
Method:                 Least Squares   F-statistic:                     24.31
Date:                Tue, 12 Sep 2023   Prob (F-statistic):           2.07e-28
Time:                        13:36:24   Log-Likelihood:                -4182.1
No. Observations:                3900   AIC:                             8378.
Df Residuals:                    3893   BIC:                             8422.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      16.1677      3.797      4.258      0.000       8.723      23.612
C(race)[T.2]   -0.1117      0.035     -3.232      0.001      -0.180      -0.044
C(race)[T.3]   -0.0820      0.045     -1.833      0.067      -0.170       0.006
workblks        0.0234      0.010      2.344      0.019       0.004       0.043
inc10k          0.0322      0.004      9.202      0.000       0.025       0.039
year           -0.0120      0.002     -5.917      0.000      -0.016      -0.008
cohort          0.0051      0.001      5.514      0.000       0.003       0.007
==============================================================================
Omnibus:                      137.717   Durbin-Watson:                   1.974
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              151.352
Skew:                          -0.478   Prob(JB):                     1.36e-33
Kurtosis:                       2.870   Cond. No.                     9.39e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.39e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

According to this OLS, for every increase in positive view of black people, people on average increase their health by 0.0234 net of all other variables. This is statistically significant and confirms my assumption that those with less bias have better health (ie those with more prejudice have worse health).

For every 10,000 dollar increase in income, people on average have a 0.0322 increase in health, net of all other variables. This is also statistically significant and confirms my assumption that more wealth leads to better health.

Black people compared to white people have 0.1117 points lower health, net of all other variables. This is also statistically significant and makes sense as structural determinents of health usually leads to black people having worse health outcomes.

People who fall into the 'other' race category compared to whites have on average 0.0820 points lower health, net of all variables. This is not quite statistically significant (p-value of 0.067) but again follows similar logic to black people as non-white individuals tend to face worse structural determinents of health leading to worse health.

Laslty, for every year increase in cohort, people on average have a 0.0051 increase in health, net of all other variales,which is statistically significant. This does not confirm previous assumptions of older individuals having worse health.

Critique of OLS

While this OLS does support my hypothesis, it may not be the best model for my question as how people perceive black people might be endogenous to health. There may be an omitted variable that is driving both health and perception of black people that is obscuring the direct relationship between racism and health. For example, it is possible that some people are just more generally positive in life and thus have a more optimisitc view of their health and kinder perceptions of people in general. This type of omitted variable is difficult to measure and thus cannot be controlled for. Instead an instrumental variable could be used to circumvent this variable by using a proxy of such for perception of black people that is not impacted by general optimism. Using a proxy that is more or less random that would still be highly correlated with percetion of black people would allow for stronger causal conclusions to be made.

Instrumental Variable

Instrument

For my isntrumental variable I am going to use 'Racwork', whichis the racial makeup of workplace. This variable ranges from 1 (all white workplace) to 5 (all blakc workplace). A value of 6 indicates that the person works alone, which for the purpose of this excercise I will exclude.

I chose this variable as it is somewhat random, a person doesn't necessarily choose the racial makeup of their workplace. It also should not be directly related to health outcomes, it shouldn't in theory matter for your health who you work with. Additionally, it is also not related to the potential omitted variable of being a more positive individual (ie z is uncorrelated with u).

This instrumental variable though could be related to the amount of prejudice a person has, as the more exposure a person has to people in a group, the less prejudiced they may become.

```
#drop value 6 which does not apply to analysis
data = data.drop(data[data['racwork'] == 6].index)

print(data['racwork'].max())
```

```
5.0
```

IV Model

```
from linearmodels.iv.model import IV2SLS

iv1 = IV2SLS.from_formula('health ~ inc10k + year + cohort + C(race) + [workblks  ~ racwork]', data=data).fit()
iv1
```

IV-2SLS Estimation Summary

Dep. Variable:	health	**R-squared:**	-68.475
Estimator:	IV-2SLS	**Adj. R-squared:**	-68.582
No. Observations:	3899	**F-statistic:**	557.81
Date:	Tue, Sep 12 2023	**P-value (F-stat)**	0.0000
Time:	14:24:55	**Distribution:**	chi2(7)
Cov. Estimator:	robust		

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
inc10k	0.1146	0.3393	0.3377	0.7356	-0.5505	0.7797
year	-0.0098	0.0205	-0.4752	0.6346	-0.0500	0.0305
cohort	0.0463	0.1698	0.2724	0.7853	-0.2866	0.3792
C(race)[T.1]	-55.116	297.14	-0.1855	0.8528	-637.51	527.27
C(race)[T.2]	-51.949	283.73	-0.1831	0.8547	-608.06	504.16
C(race)[T.3]	-56.205	301.27	-0.1866	0.8520	-646.69	534.28
workblks	-5.2222	21.608	-0.2417	0.8090	-47.573	37.129

Endogenous: workblks
Instruments: racwork
Robust Covariance (Heteroskedastic)
Debiased: False
id: 0x7e62713f2ce0

According to this IV model, for every increase in perception of black people, people on average decrease their overall health scores by 5.2222, net of all other variables. This is no longer statistically significant in this model and is in the reverse direction as the OLS model. This is likely due to racwork not being a valid instrument, which will be discussed more in the diagnostic section. The larger standard error of 21.608 in this model compared to 0.010 of the OLS mdoel though is expected, as IV our less efficient and increases standard errors and lowers p-values.

Conceptually, racwork being a poor IV becomes clear on reflection. If my assumption is that cortisol increases when a person is exposed to someone who is a part of the group they are biased against, exposure in the workplace would thus have the effect of raising cortisol and worsening health (assuming they are indeed racist to begin with). Since the composition of the workplace is later shown to not be strongly correlated with my indicator of prejudice, it is only adding random noise to my model, producing hard to interpret coefficients as opposed to causal relationships.

The control variables in this model were also highly impacted by the inefficient nature of IV models as none of these coefficients are statistically significant anymore. For income, the direction and magnitude was relatively strong however, with a change from 0.0322 in the OLS model to a coefficient of 0.1146 in the IV model. Similarly for the cohort coefficient, it is fairly similar between the two models with OLS being 0.0051 and the IV model being 0.0463. This may indicate that the relationship between age/income and health is rather robust.

The coefficient for the race controls however were wildly impacted by the IV model, going from modest levels of -0.1117 for black individuals and -0.0820 for 'other' individuals to -51.949 and -56.205 respectively. This is another indicator that the instrumental variable of racwork is introducing a lot of likely invalid noise into the model.

Diagnostics

```
#determine if racwork is highly correlated with independent variable

lm_strong = smf.ols('workblks  ~ racwork', data = data).fit()

# Print the model summary
print(lm_strong.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               workblks   R-squared:                       0.006
Model:                            OLS   Adj. R-squared:                  0.006
Method:                 Least Squares   F-statistic:                     22.56
Date:                Tue, 12 Sep 2023   Prob (F-statistic):           2.11e-06
Time:                        14:24:41   Log-Likelihood:                -6109.6
No. Observations:                3899   AIC:                         1.222e+04
Df Residuals:                    3897   BIC:                         1.224e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
```

```
Intercept      2.5966      0.048     53.631      0.000       2.502       2.691
racwork        0.1030      0.022      4.750      0.000       0.060       0.145
==============================================================================
Omnibus:                       52.398   Durbin-Watson:                   1.883
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               93.484
Skew:                           0.040   Prob(JB):                     5.01e-21
Kurtosis:                       3.754   Cond. No.                         6.84
==============================================================================

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

The first verifiable assumption of IV is that your instrument is strongly correlated with your independent variable, which is indicated by a high r-squared when regressing the IV on the independent variable. For my chosen IV, this assumption is violated with a very weak r-squared of 0.006. This indicated that there is very little relevance to my variable and it should not be used as an instrumental variable for my variables of interest.

```
#determine if racwork is highly correlated with dependent variable on its own

lm_pred = smf.ols('health  ~ racwork', data = data).fit()

# Print the model summary
print(lm_pred.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 health   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.004
Method:                 Least Squares   F-statistic:                     16.90
Date:                Tue, 12 Sep 2023   Prob (F-statistic):           4.03e-05
Time:                        15:42:45   Log-Likelihood:                -4243.5
No. Observations:                3899   AIC:                             8491.
Df Residuals:                    3897   BIC:                             8504.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.2775      0.030     75.916      0.000       2.219       2.336
racwork       -0.0552      0.013     -4.111      0.000      -0.082      -0.029
==============================================================================
Omnibus:                      149.568   Durbin-Watson:                   1.946
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              165.517
Skew:                          -0.499   Prob(JB):                     1.14e-36
Kurtosis:                       2.854   Cond. No.                         6.84
==============================================================================

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

Another assumption of the IV model is that the instrument is not directly correlated with the dependent variable, as a quick check I regressed my instrument on the dependent variable to see if there is a relationship between them without my independent variable as an intermediate. This showed that for every increase in diversity in the workplace, a person on average has a decrease in health by -0.0552, which is statistically significant. As there seems to be a relationship between these two variables by themselves, it seems that this assumption is violated and again my chosen instrument is not valid.

```
# Tests for IV specification
# Wu-Hausman test of exogeneity
print(iv1.wu_hausman())

# First Stage Diagnostics
print(iv1.first_stage)
```

```
Wu-Hausman test of exogeneity
H0: All endogenous variables are exogenous
Statistic: 4.8564
P-value: 0.0276
Distributed: F(1,3891)
    First Stage Estimation Results
=====================================
                            workblks
-------------------------------------
R-squared                     0.0455
Partial R-squared          1.753e-05
```

```
Shea's R-squared             1.753e-05
Partial F-statistic             0.0602
P-value (Partial F-stat)        0.8062
Partial F-stat Distn           chi2(1)
========================== ===========
inc10k                          0.0157
                               (3.0261)
year                            0.0004
                               (0.1153)
cohort                          0.0078
                               (5.1496)
C(race)[T.1]                   -13.490
                              (-2.1639)
C(race)[T.2]                   -12.871
                              (-2.0649)
C(race)[T.3]                   -13.684
                              (-2.1921)
racwork                         0.0061
                               (0.2453)
--------------------------------------

T-stats reported in parentheses
T-stats use same covariance type as original model
```

Again, these diagnostic tests indicate that my chosen instrumental variable was a weak instrument, with an F statistic of 0.0602, which is much much smaller than the general rule of thumb of 10 or higher.\

However, the Wu-Hausman test had a t-statistics of 4.8564, which was statistically significant, which does indicate that my independent variable was endogenous to health. This suggests that an instrumental variable is needed to improve the first OLS model.

Conclusion

Overall, my chosen instrumental variable was weak and was not a valid choice to model the relationship between racism and health. However, the Wu-Hausman test indicated that there was an omitted variable that should be addressed, meaning my OLS is also not a good model to describe the relationship between my variables of interest. As such, there is no concrete conclusion from these models and a new instrumntal variable should be found to try to get a clearer picture of how racism and health are related. While the OLS did support my initial hypothesis, it is also possible that my proxy for racism (how individuals view the hard-working nature of black people) could also not be a valid proxy. More exploration should be done finding other potential indicators of racism and also crucially other instrumental variables that are strong are necessary to really answer my question of interest. There does appear to be a strong relationship between income and age on health and these variables should definitely be included as controls in future models.