FEM 11087 - Applied Microeconometrics

Assignment 1: Empirical Analysis

Regression Analysis with Cross-Sectional Data, Endogeneity and Instrumental Variable Estimation

Group 23

Kees-Piet Barnhoorn 729384
Tyler McGee 781846
Andres Pinon 775387
Jolien Schaeffers 779155

16 September 2025

Question 1 [0.6 points]

First generate the variable BMI, where BMI equals weight in kg divided by height in meters squared (BMI = weight/(height^2)). Construct a categorical variable for BMI that considers the commonly used categories: i) underweight, BMI below 18.5; ii) normal weight, BMI larger or equal to 18.5 and lower than 25; iii) overweight, BMI larger or equal to 25 and lower than 30; iv) obese, BMI of 30 or higher. Compute and report the prevalence of overweight and obesity by ethnic group (black vs non-black). What differences do you observe?

We begin by creating the `bmi` variable. Note that, since Body Mass Index[1] is calculated as

$$\text{BMI} = \frac{\text{weight (kg)}}{\text{height (m)}^2} \tag{1.1}$$

we must first convert our `height` variable from centimeters to meters using a conversion factor of 100. We can generate both `height_m` and `bmi` through the use of the `gen` command:

[1] World Health Organization. *Obesity: preventing and managing the global epidemic: report of a WHO consultation*. WHO technical report series 894. Geneva: World Health Organization, 2000

```
gen height_m = height / 100
gen bmi = weight / (height_m^2)
```

Our categorical BMI variable `bmi_cat` can then be constructed through the use of the **`replace`** command:

```
gen bmi_cat = .
replace bmi_cat = 1 if bmi < 18.5 & !missing(bmi)
replace bmi_cat = 2 if bmi >= 18.5 & bmi < 25 & !missing(bmi)
replace bmi_cat = 3 if bmi >= 25 & bmi < 30 & !missing(bmi)
replace bmi_cat = 4 if bmi >= 30 & !missing(bmi)
```

The addition of `!missing(bmi)` to the **`if`** statements prevents Stata from assigning observations with missing `bmi` values a `bmi_cat` value of 4. This occurs due to Stata coding all missing values (., .a, .b, .c, ..., .z) as larger than any non-missing value.[2]

*a) Compute and report the **prevalence of overweight and obesity** by ethnic group (black vs non-black). What differences do you observe?*

Binary indicators for overweight and obesity status are generated:

```
gen overweight = (bmi_cat >= 3) if !missing(bmi_cat)
gen obese = (bmi_cat == 4) if !missing(bmi_cat)
```

Importantly, our overweight indicator includes both overweight (BMI 25-29.9) and obese (BMI $\geq$ 30) individuals, allowing us to compare the prevalence of both categories by ethnic group. Using **`tab`**:

```
tab black overweight, row missing
tab black obese, row missing
```

[2] William Gould. *Logical expressions and missing values*. Stata FAQ. Stata Corp. URL: https://www.stata.com/support/faqs/data-management/logical-expressions-and-missing-values/ (visited on 09/14/2025)

```
    Race: |
    Black |
  (1=Yes, |      overweight
    0=No) |         0          1 |     Total
----------+----------------------+----------
        0 |        87        296 |       383
          |     22.72      77.28 |    100.00
----------+----------------------+----------
        1 |        12         50 |        62
          |     19.35      80.65 |    100.00
----------+----------------------+----------
        . |         1          0 |         1
          |    100.00       0.00 |    100.00
----------+----------------------+----------
    Total |       100        346 |       446
          |     22.42      77.58 |    100.00
```

```
    Race: |
    Black |
  (1=Yes, |         obese
    0=No) |         0          1 |     Total
----------+----------------------+----------
        0 |       225        158 |       383
          |     58.75      41.25 |    100.00
----------+----------------------+----------
        1 |        27         35 |        62
          |     43.55      56.45 |    100.00
----------+----------------------+----------
        . |         1          0 |         1
          |    100.00       0.00 |    100.00
----------+----------------------+----------
    Total |       253        193 |       446
          |     56.73      43.27 |    100.00

```

As seen in the output, the rate of overweight is similarly high amongst ethnic groups (75.1% vs 76.9%); however, obesity rates are markedly higher in black individuals (53.8% vs 40.1%). This suggests that, while the total overweight rates are similar, black overweight individuals are more likely to fall into the obese category.

*b) Make an appropriate graph to **compare income distributions** across ethnic groups and discuss what you see.*

A clear comparison of income can be appreciated through the use of a box plot:

```
graph box income, over(black)
```

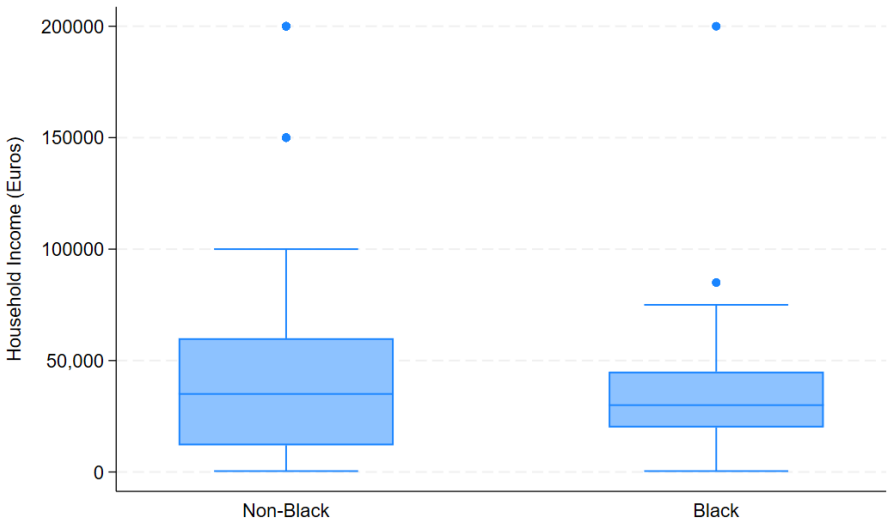

Figure 1.1: Income Distribution by Ethnic Group

Looking at figure 1.1, we can see that both ethnic groups have a similar median income; however, the non-black group presents a larger interquartile range (IQR) than the black group, meaning higher income variability. It is also worth noting that the upper whiskers show how, within the normal income range (excluding outliers), the non-black group reaches substantially higher income levels.

*c) If there are **missing values** for BMI, discuss how they may impact the validity of your regression analysis.*

The missing BMI values in our sample effectively reduce the sample size, leading to a loss in the precision of our estimates. If said missingness is not random, but is instead related to other variables in our model, it would also lead to selection bias in the regression estimates.

```
gen bmi_miss = missing(bmi)
tab black bmi_miss, row missing
```

```
     Race: |
     Black |
   (1=Yes, |  bmi_miss
     0=No) |         0 |     Total
-----------+-----------+----------
         0 |       383 |       383
           |    100.00 |    100.00
-----------+-----------+----------
         1 |        62 |        62
           |    100.00 |    100.00
-----------+-----------+----------
         . |         1 |         1
           |    100.00 |    100.00
-----------+-----------+----------
     Total |       446 |       446
           |    100.00 |    100.00
```

The above table show how BMI missingness is not related to the `black` variable. In order to check if its related to income, we construct a categorical income variable and compute its missingness:

```
gen income_cat = .
replace income_cat = 1 if income < 25000 & !missing(income)
replace income_cat = 2 if income >= 25000 & income < 50000 & !missing(income)
replace income_cat = 3 if income >= 50000 & income < 100000 & !missing(income)
replace income_cat = 4 if income >= 100000 & !missing(income)
tab income_cat bmi_miss, row missing
```

```
            |       bmi_miss
income_cat  |         0          1 |     Total
------------+----------------------+----------
          1 |       165          6 |       171
            |     96.49       3.51 |    100.00
------------+----------------------+----------
          2 |       132          3 |       135
            |     97.78       2.22 |    100.00
------------+----------------------+----------
          3 |       116          3 |       119
            |     97.48       2.52 |    100.00
------------+----------------------+----------
          4 |        34          1 |        35
            |     97.14       2.86 |    100.00
------------+----------------------+----------
      Total |       447         13 |       460
            |     97.17       2.83 |    100.00
```

*d) If there are **unreliable or implausible** values in any of the variables of interest, describe how you would identify them and drop them before proceeding with the rest of the analysis.*

Unreliable and implausible values can be identified by examining the data with the help of **`summarize`** and **`list`**.

```
sum income height weight bmi
list income height weight if income < 0
list height weight if height < 0 | height > 272
list bmi if bmi < 10 | bmi > 60
```

Programmatically, `flag_` variables can be generated when specific thresholds are passed, in which case they can be dropped with the **`drop`** command.

```
gen flag_income = (income < 0)
gen flag_height = (height < 0 | height > 250)
gen flag_bmi = (bmi < 10 | bmi > 60)
drop if flag_income == 1 | flag_height == 1 | flag_bmi == 1
```

In our case, we set the lower income threshold at 0, since it is not plausible to earn a negative income. For height, we chose 272cm as the upper height threshold, the highest reported height ever recorded in a human, as well as a lower threshold of 0, since it is not possible to stand at a negative height. For BMI, we set a lower threshold of 10, following the same reasoning as with height (lowest BMI ever reported), as well an upper threshold of 60.

Question 2 [0.9 points]

Estimate a multivariate regression model explaining income as a function of BMI and whether the individual is black. Thus, estimate:

$$Income = \beta_0 + \beta_1 BMI + \beta_2 Black + u$$

Using `reg` with robust SE estimators:

```
reg income bmi black, robust
```

Table 2.1: Linear regression

	Coef.	Robust SE	p-value	[95% Conf.	Interval]
bmi	-307.3532	277.6654	0.2689	-853.0618	238.3554
black	-4839.7565	5108.1110	0.3439	-14878.9600	5199.4471
cons	51459.4549	8992.8883	0.0000***	33785.3216	69133.5882
Number of obs.	445				
F	1.0667				
Prob > F	0.3450				
R-squared	0.0047				
Adj R-squared	0.0002				
Root MSE	39216.1566				

*a) Interpret the estimated coefficients of all the explanatory variables (**sign, magnitude, and significance**).*

For the explanatory variable `bmi`, the negative coefficient implies a negative relationship. The magnitude of the coefficient, 307.3532, implies that for each unit increase in BMI, an individual's observed annual income would decrease by roughly

307$, ceteris paribus. Finally, it can be determined that the `bmi` variable is not significant, as its p-value (0.269) exceeds our significance level (0.05).

In the case of the variable `black`, the negative coefficient implies an also negative relationship with the dependent variable. The magnitude of the coefficient, 4839.756, implies a decrease of roughly 4840$ in observed annual income if an individual belongs to the black ethnic group, ceteris paribus. As in the case of `bmi`, `black`'s p-value (0.344) exceeds the significance level, meaning it is not significant.

*b) What is the estimated 95% Confidence Interval of β_1? What can you conclude based on the information in this confidence interval about the **effect of BMI on income**?*

As seen in the regression above, the estimated 95% CI of β_1 is [-788.01, 286.60]. Since 0 falls within it, we cannot reject the null hypothesis that BMI has no effect on income at the 5% significance level. This aligns with the conclusions derived from analyzing the p-values in the previous question.

*c) What is the **relative magnitude** of the effect of being black on income? Interpret the relative magnitude.*

The relative magnitude of being black on income can be explained as the % change in income experienced if an individual belongs to the black ethnic group, everything else notwithstanding. We can calculate it as:

$$\text{Relative Magnitude} = \frac{-4,839.75}{41,514.35} = -0.117 \ (11.7\%) \tag{2.1}$$

We can therefore determine that the coefficient for black ethnicity represents roughly 11.65% of mean annual income.

Question 3 [0.9 points]

In the previous question, we have assumed that the association between income and BMI is linear.

a) Do you think this assumption is likely to hold? Explain.

The above assumption is not likely to hold. In reality, we would expect values both at the very high and very low BMI thresholds to be associated with lower income, suggesting a non-linear relationship.

Extremely low BMI could account for malnutrition or severe health issues, both of which we would expect to be related to lower incomes, as malnutrition shows difficultty accessing resources and severe health issues might limit an individual's capacity to earn a stable income.

While it is true that a higher BMI might be related to abundance, an extremely high BMI would be related to medical problems as severe as those found in extremely low BMI cases. It is worth noting that social stigma could also negatively impact an individual's ability to secure an income while extremely overweight.

b) Add $\boldsymbol{BMI}^2$ *to the regression of question 2 and estimate it. What is the estimated effect of BMI on income? In your answer, interpret the effect at two different points of the BMI distribution.*

```
gen bmi2 = bmi^2
reg income bmi bmi2 black, robust
```

Table 3.1: Linear Regression: Non-linearities

	Coef.	Robust SE	p-value	[[95% Conf.	Interval]
bmi	390.9600	2387.6819	0.8700	-4301.6893	5083.6094
bmi2	-10.8482	35.0675	0.7572	-79.7683	58.0719
black	-4812.9266	5110.7583	0.3468	-14857.3955	5231.5422
cons	40704.0961	39289.7653	0.3008	-36514.3517	117922.5438
Number of obs.	445				
F	0.9913				
Prob > F	0.3967				
R-squared	0.0049				
Adj R-squared	-0.0019				
Root MSE	39256.7890				

Analyzing the new model, we can see that the `bmi` coefficient is now positive, while the `bmi2` coefficient is negative. This suggests quadratic relationship where income first increases with BMI, then decreases. It follows the form

$$\text{Income} = 40,704 + 390.96 \times \text{BMI} - 10.85 \times \text{BMI}^2 \tag{3.1}$$

We will compare the effects of BMI on income taking two points, BMI = 20 and BMI = 30. In the first case

$$40,704 + 390.96 \times 20 - 10.85 \times 20 = 44,183.2 \tag{3.2}$$

while, in the second case

$$40,704 + 390.96 \times 30 - 10.85 \times 30 = 42,667.8 \tag{3.3}$$

We see that income is lower at 30 BMI compared to 20 BMI. This coincides with our prediction that the relationship between income and BMI follows a non-linear pattern.

c) How does adding BMI^2 capture non-linearities in the relationship between BMI and income?,

As seen in the previous question, the marginal effect of BMI on income at different points of the BMI scale can have not just different magnitudes, but even different signs altogether. Since a linear model assumes a constant effect and cannot therefore accurately represent this pattern, adding BMI^2 allows the slope of our model to vary at different BMI values.

*d) What is your **preferred specification** (2 or 3)? Explain.*

The non-linear specification is preffered over the original especification, since it allows the model to capture the relationship between income and BMI without assuming a linear pattern between the variables.

Question 4 [0.7 points]

Use the log of income as the dependent variable. Start by creating this variable.

To create the `ln_income` variable, we use the **`gen`** command as follows:

```
1 gen ln_income = ln(income + 1)
```

`ln(income+1)` is the equation to ensure that data entries with income values of 0 are still included within the regression model.

a) Provide one reason why a logarithmic transformation of income may be useful in (linear) regression analysis.

A logarithmic transformation proves useful in addressing the right-skewedness of income, which arises from the fact that income's lower bound is set at 0. Additionally, it allows coefficients to be interpreted as percentage changes rather than absolute units, which is quite useful for the purposes of our analysis.

b) Estimate a regression model using OLS explaining the ***log of income*** *as a function of* ***BMI as categorical variable*** *and whether the individual is black.*

We can estimate said model using `reg`. Note that the `i.` prefix is used to treat a variable as categorical in the context of a regression. By doing so, the model will separately estimate coefficients for each of the `bmi_cat` categories; otherwise, we would be assuming a constant effect within between each category.

Adding `b2` to the prefix specifies that `bmi_cat2` (Normal Weight) is omitted in order to prevent multicollinearity. `bmi_cat2` has been chosen due to the fact that, while the results are statistically equivalent independently of which category is omitted, in the case of BMI Normal Weight is a better baseline than Underweight.

```
1 reg ln_income ib2.bmi_cat black, robust
```

Table 4.1: Linear Regression: Categorical BMI

	Coef.	Robust SE	p-value	[95% Conf.	Interval]
bmi_cat 1	0.1791	0.1496	0.2317	-0.1148	0.4731
bmi_cat 3	0.1807	0.1857	0.3313	-0.1844	0.5457
bmi_cat 4	0.1574	0.1754	0.3702	-0.1874	0.5022
black	-0.0464	0.1729	0.7885	-0.3862	0.2934
cons	9.9475	0.1496	0.0000***	9.6536	10.2415
Number of obs.	445				
F	.				
Prob > F	.				
R-squared	0.0029				
Adj R-squared	-0.0061				
Root MSE	1.3332				

*c) Interpret the estimated coefficients of all the explanatory variables (**sign, magnitude, and significance**).*

The estimated coefficients across the different BMI categories all show a positive relationship between BMI and income. They are also similar in terms of magnitude, with categories 1, 3 and 4 having magnitudes of 0.1791, 0.1807 and 0.1574, respectively. In terms of statistical significance, all three categories have corresponding p-values well above the 0.05 significance level, meaning we reject them as statistically significant.

In the case of `black`, its coefficient is estimated at -0.0464. Transforming the coefficient using the formula

$$\% \text{ change} = (e^{\beta} - 1) \times 100 \tag{4.1}$$

meaning that belonging to the black ethnic group is associated with a decrease of roughly 4.75% in annual income, all other values notwithstanding. It's statistical significance is rejected, since its p-value, 0.789, is well above the significance level.

*d) Extend the model to estimate if the relationship between **BMI (in categories)** and income is different **across ethnic groups**. What do you conclude?*

To estimate the differences in the relationship between `bmi_cat` and income across different ethnic groups, we can use interaction variables. Conceptually, the model would be specified as

$$\begin{aligned} \ln(\text{Income}) = {} & \beta_0 + \beta_1 \text{Underweight} + \beta_2 \text{Overweight} + \beta_3 \text{Obese} + \beta_4 \text{Black} \\ & + \beta_5 (\text{Underweight} \times \text{Black}) + \beta_6 (\text{Overweight} \times \text{Black}) \\ & + \beta_7 (\text{Obese} \times \text{Black}) + u \end{aligned} \tag{4.2}$$

In Stata, that can be easily achieved through the use of the `####` operator.

```
reg income ib2.bmi_cat##i.black
```

Table 4.2: Linear Regression: Interaction Effects

	Coef.	Robust SE	p-value	[95% Conf.	Interval]
bmi_cat 1	0.0621	0.1557	0.6899	-0.2438	0.3681
bmi_cat 3	0.0797	0.1938	0.6812	-0.3013	0.4606
bmi_cat 4	-0.0373	0.1870	0.8421	-0.4048	0.3302
black	-1.0019	0.4560	0.0285*	-1.8980	-0.1057
bmi_cat 1 · black	0.0000	.	.	0.0000	0.0000
bmi_cat 3 · black	0.7923	0.6054	0.1913	-0.3976	1.9821
bmi_cat 4 · black	1.3837	0.4839	0.0044**	0.4327	2.3347
cons	10.0645	0.1557	0.0000***	9.7586	10.3705
Number of obs.	445				
F	.				
Prob > F	.				
R-squared	0.0223				
Adj R-squared	0.0090				
Root MSE	1.3231				

The extended model estimates clear differences in the relationship between BMI and income across both ethnic groups. While for non-black individuals BMI categories have small and statistically non-significant effects on income, for black individuals in both the Overweight and the Obese categories there is strong relationship between BMI and income.

The coefficient for `bmi_cat3 * black`, 0.7923, implies that being both black and overweight is related to an increase of 139.17% in annual income, compared to just a 8.3% increase for non-black, overweight individuals. The p-value for this interaction term is given at 0.146, which could be considered marginally significant.

In the case of `bmi_cat4 * black`, 1.3837, implies an even bigger effect in cases in which an individual is both black and obese, in which case there would be an increase of 284.36% in annual income over the baseline compared to a small decrease of 3.8% for non-black, obese individuals. The p-value for this interaction is given at 0.004, well below our significance level of 0.05. This implies that coefficient $\beta_{\text{obese}\times\text{black}}$ is statistically quite significant.

Question 5 [0.5 points]

Are any of the models in Question 2, Question 3 and Question 4 ***correctly specified****? Explain with your own words. (Note: For this question, focus on functional form).*

In order to test for omitted variables in our models, we can run Ramsey RESET tests on the regression models from questions 2, 3 and 4. In Stata, that can be comfortably done using the `estat ovtest` command:

```
reg income bmi black, robust
(output omitted)

estat ovtest

Ramsey RESET test for omitted variables
Omitted: Powers of fitted values of income

H0: Model has no omitted variables

F(3, 439) =   2.63
 Prob > F = 0.0496
```

```
reg income bmi bmi2 black, robust
(output omitted)

estat ovtest

Ramsey RESET test for omitted variables
Omitted: Powers of fitted values of income

H0: Model has no omitted variables

F(3, 438) =   3.24
 Prob > F = 0.0219
```

```
reg ln_income ib2.bmi_cat black, robust
(output omitted)

estat ovtest

Ramsey RESET test for omitted variables
Omitted: Powers of fitted values of ln_income

H0: Model has no omitted variables

F(2, 438) =    4.35
 Prob > F = 0.0135
```

```
reg ln_income ib2.bmi_cat##i.black
(output omitted)

estat ovtest

Ramsey RESET test for omitted variables
Omitted: Powers of fitted values of ln_income

H0: Model has no omitted variables

F(2, 438) =    4.35
 Prob > F = 0.0135
```

Perhaps unintuitively, a p-value <= 0.05 is, in the context of this test, *not* a promising result when analyzing a model. The RESET test is specified as:

$$H_0 : \text{The model is correctly specified (no omitted variables)}$$
$$H_1 : \text{The model is misspecified (omitted variables exist)}$$

Looking at the RESET results for our models, the fact that all p-values fall below the significance level implies that, actually, we can reject all the null hypotheses and conclude that *none* of our models so far have been correctly specified!

Question 6 [0.9 points]

*a) Reflect on the interpretation of the BMI coefficient. Does it capture the **causal effect** of body fat on income?*

The BMI coefficient reflects the relation between BMI and income; however, it does not capture the causal effect of body fat on income. This means that we cannot confidently establish causality between the two variables by looking at the regression alone. Many scenarios, such as reverse causality, bidirectional causality or even no true causal relationship confounded by omitted variables could explain the coefficient estimated by our model (which is, worth noting, not statistically significant in the regression models discussed so far).

*b) Draw a **DAG** illustrating how an **omitted variable** could bias the estimated effect of BMI on income and indicate one such variable.*

A possible omitted variable would be educational attainment, due to its strong effect on both income and BMI. Higher education could likely be positively correlated to income and, through factors such as health knowledge, organizational skills and discipline, negatively correlated to BMI scores. The following DAG illustrates this dynamic:

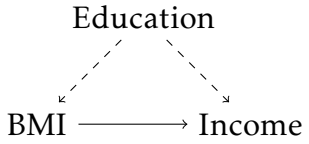

Figure 6.1: DAG showing omitted variable bias

*c) Based on your DAG, would the resulting bias in the BMI coefficient be **upward or downward**? Explain.*

Since Education would likely be positively correlated to the dependent variable Income and negatively correlated to the explicative variable BMI, we would expect a model that omits education to be biased towards showing a negative relationship between income and BMI, even if the true effect were to be neutral or positive. In our model, the BMI variables have been trying to account for both their own effect and the effect of education on income.

*d) Identify a variable (not necessarily observed in your dataset) that could act as a **collider** in this context. Use a DAG to illustrate and explain why conditioning on it would lead to bias.*

A collider variable is a variable caused by both the dependent and explicative variables. In the case of our model, a collider variable could be Health Insurance, since both higher BMI and lower income are likely correlated to worse health coverage. A simple DAG illustrating this effect would look as follows:

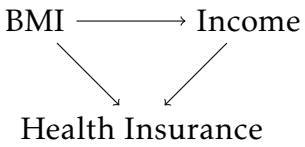

Figure 6.2: DAG showing Health Insurance as a collider

Question 7 [0.6 points]

*Do you think that **reverse causality** influences the OLS estimate of BMI? Explain.*

In the case of income and body fat, reverse causality most likely exists due to the fact that income can influence body fat through differences in access to (healthy) food, exercise and healthcare. This introduces endogeneity bias into the model, which violates the OLS assumption that the explanatory variable (in our case, BMI) is uncorrelated to the error.

In our example, assuming that higher income is indeed related to an overall healthier lifestyle, part of what the model estimates as the impact of BMI on income would instead be the reverse effect of income differences on BMI.

*Estimate a multivariate regression model explaining **BMI as a function of income** and whether the individual is black.*

A model utilizing income and black as explicative variables for BMI can be estimated with:

```
reg bmi income black, robust
```

Table 7.1: Linear Regression: Reverse Causality

	Coef.	Robust SE	p-value	[95% Conf.	Interval]
income	-0.0000	0.0000	0.2507	-0.0000	0.0000
black	1.7927	0.9341	0.0556	-0.0431	3.6286
cons	29.9964	0.4493	0.0000***	29.1133	30.8795
Number of obs.	445				
F	2.5861				
Prob > F	0.0765				
R-squared	0.0126				
Adj R-squared	0.0081				
Root MSE	6.2808				

Is the coefficient of income statistically significant? Does this provide evidence that the variable BMI suffers/does not suffer from reverse causality in the previous models? Discuss.

Based on the above results, the income coefficient is definetely not statistically significant. It's p-value is estimated at 0.251, meaning that we cannot reject the null hypothesis that income has no effect on BMI; this, however, does not provide evidence about whether or not the dependent variable suffers from reverse causality.

Had the income coefficient been significant, it would still not constitute proof of causality, but it could have served as a hint about the nature of the underlying relation between the variables in our model.

Question 8 [1 point]

We now consider using a variable that indicates how many times in the past seven days an individual has had sweetened drinks as an instrument for BMI.

Estimate the model of Question 2 by 2SLS, using `drinks` *as an instrumental variable for BMI.*

We can estimate models by 2SLS using the `ivregress 2sls` command:

```
ivregress 2sls income (bmi = drinks) black, robust first
```

Table 8.1: 2SLS: First-stage Regression

	Coef.	Robust SE	p-value	[95% Conf.	Interval]
black	1.8653	0.9142	0.0419*	0.0685	3.6620
drinks	-0.0631	0.0371	0.0891	-0.1360	0.0097
cons	29.9576	0.3832	0.0000***	29.2044	30.7107
Number of obs.	445				
F	3.7073				
Prob > F	0.0253				
R-squared	0.0149				
Adj R-squared	0.0104				
Root MSE	6.2735				

Table 8.2: 2SLS: IV Regression

	Coef.	Robust SE	p-value	[95% Conf.	Interval]
bmi	8141.4777	5978.4412	0.1733	-3576.0517	19859.0072
black	-20346.0960	13430.1982	0.1298	-46668.8007	5976.6087
cons	-199154.8722	176930.3905	0.2603	-545932.0653	147622.3210
Observations	445				
Wald χ^2	2.36				
Prob > χ^2	0.3069				
Root MSE	65812.339				

a) Write down the estimated first stage of the model.

The first stage of the model would be specified as

$$\begin{aligned} \text{BMI} &= \pi_0 + \pi_1 \text{drinks} + \pi_2 \text{black} + v \\ \text{BMI} &= 29.958 + 1.865 \cdot \text{drinks} + 0.631 \cdot \text{black} + v \end{aligned} \tag{8.1}$$

*b) **Interpret** the estimated coefficient for `drinks` in the first stage regression. Is the estimated coefficient of BMI and its significance obtained with the IV estimator different from the OLS estimator in question 2? Does this suggest that the variable is endogenous or not? Discuss [Note: For this question, do not perform any additional analysis]*

Drinking one sweet drink decreases the BMI value by 0.06 points on average. This coefficient is significant at the 10% significance level; however it is important to check the strength of the instrument. Here we see that its coefficient magnitude is rather low in addition to the fact that BMI does not have a high value range.

We see that the F value computed from the F-test is 3.7073. As a rule of thumb, this can be interpreted as a weak instrument, since the value of F is lower than 10 ($3.7073 < 10$). Additionally, we look at the difference between the coefficient of BMI and its significance between this IV regression, and the OLS regression in question 2. In question 2, the regression model states that BMI has a coefficient of -307.353, with an insignificant p-value of 0.269 (which is insignificant at the 10% significance level).

In the IV regression, BMI has a coefficient of 8141.478 and an insignificant p-value of 0.173 (which is insignificant at the 10% significance level). We see a huge difference in the coefficient's magnitude, as well as sign. Despite the statistical insignificance, this difference in coefficient magnitude and sign suggests that there is a potential omitted variable bias which underestimated the coefficient in the first model.

*c) Perform a **formal test** of the null hypothesis that BMI is **exogenous**. What do you conclude? Explain.*

To perform a formal test on the null hypothesis that BMI is exogenous, we perform an estat test within stata with the following code:

```
estat endogenous

Tests of endogeneity
```

```
H0: Variables are exogenous

Robust score chi2(1)            =  5.02425  (p = 0.0250)
Robust regression F(1,441)      =  5.61163  (p = 0.0183)
```

This formal test shows through the p-value of the F-test that the null hypothesis of variables being exogenous; we have enough evidence to reject the null hypothesis at the 5% significance level, but not the 1% significance level ($0.01 < 0.018 < 0.05$). This is a clear sign that the variables are endogenous which means there is an apparent issue of endogeneity within the sample size which is also supported by the difference in the coefficients.

Question 9 [0.7 points]

*a) Draw a **DAG** that illustrates the assumptions required for `drinks` to be a good instrument for BMI.*

b) Explain these assumptions in your own words.

The four assumptions that are required for drinks to be a good instrument are as follows:

- Omitted Variable Bias: Unobserved factors exist which could influence both the dependent and explanatory variable in the model.
- Relevance: A relevant instrument is one in which it is correlated with the endogenous variable such as o Cov(x,z) != 0, where x is the endogenous variable and z is the instrumental variable.
- Validity: A valid instrumental variable is also one that is uncorrelated with the unobserved factors of the model so that o Cov(u,z) = 0, where u is the unobserved variables and z is the instrumental variable.
- Additionally, the instrumental variable must only affect the dependent variable through the endogenous variable, it is not allowed to directly affect the dependent variable.

c) In your opinion, do these assumptions hold? Discuss without any further analysis. You may use evidence from previous questions.

With regards to the analyses, we've previously conducted; we can safely say that we believe that the omitted variable bias does hold. This is supported by the RESET tests conducted in question 5 which proved that models already rejected the null hypothesis that there were no omitted variables. We believe that the additional instrumental variable in drinks is not the only omitted variable that has an effect, and there could be additional variables such as health status that are not included in the model which affect our dependent and independent variables.

For the relevance assumption: we agree that the instrumental variable drinks could very well be influencing BMI scores, considering the regression outputs we conducted and from the health hazards that consuming large amounts of sugar can have. We do not agree however, with the claim that it is a strong instrument, as drinks are only one item that causes BMI levels to change; and there are so many other food items and activities that play a huge factor in determining the level of BMI of an individual.

For the validity assumption: We believe that this assumption would not hold up for the instrumental variable drinks as there are multiple different ways that the consumption of sweet drinks influences the level of income in an individual. An example of this would stem from level of productivity. Sweet drinks (that contain sugar) are commonly used to feel more energized which allows individuals to get more work done;this work productivity can definitely have an effect on the level of annual income as more productivity individuals are believed to have higher incomes than those who are not as productive.

Question 10 [0.3 points]

What is your preferred model? Why?

Statistically speaking, we cannot say which of the models computed are our preferred model. The regression models throughout this assignment have all resulted in significantly low R-Squared values, which indicate that very little of the variation in the relationships examined are explained. In terms of statistical significance, we'd prefer the IV 2SLS regression as the first stage regressions explain the endogenous variable of BMI very well with significant p-values of our black and drinks variable at the 5% and 10% level respectively. However, there is also an argument for the OLS regression as it is efficient and also the most unbiased indicator, which would make sense to use as the tests showed insignificant evidence of

omitted variable bias.

To conclude, although the IV regression has the best statistical significance (even though all models are mostly insignificant), we believe that the regression model from 4.d, which examines the percentage change in income as a function of the effect of BMI across the different categories of BMI and whether the individual is black. We feel that this regression model does the best job at describing how income changes in relation to BMI as well as race, by categorizing BMI which interacts with the race of the individual to showcase the change in the effect of BMI on income.

Question 11 [0.4 points]

Researchers want to estimate the effect of ***peers' risky health behaviors*** *(alcohol consumption) on the* ***academic performance of adolescents*** *by exploiting the quasi-exogenous assignment of high-school students across classes. They want to use the* ***peers' fathers' drinking behavior*** *as an instrumental variable for peers' risky health behaviors. They estimate a linear regression model with peers' risky health behaviors as the dependent variable (measured as the average number of times peers consumed alcohol in the past month) and the peers' fathers' drinking as an explanatory variable. The coefficient of this explanatory variable is* ***positive and statistically significant****.*

Based on this information, what is your ***assessment*** *of this identification strategy to estimate the effect of peers' risky health behaviors on the academic performance of adolescents?*

The first-stage regression results indicate that peers' fathers' drinking behavior is positively and significantly correlated with peers' risky alcohol consumption, supporting the relevance criterion for a valid instrument. This suggests that variation in peers' fathers' drinking can explain variation in peers' risky behaviors, which is essential for identifying the causal effect of interest.

However, the validity of the instrumental variable critically hinges on the exclusion restriction, which requires that peers' fathers' drinking affects adolescents' academic performance only through peers' risky health behaviors and not via any other direct or indirect pathways. This assumption is difficult to verify and potentially problematic. Peers' fathers' drinking may be correlated with broader family or community characteristics, such as socioeconomic status, parental involvement, or cultural norms, that could independently influence academic outcomes. Without convincing evidence that these confounding channels are ruled out or adequately controlled for, the exclusion restriction remains questionable.

Finally, the success of the identification strategy also depends on the quasi-exogeneity of student assignment across classes. If the assignment process is truly random or as good as random, it strengthens the plausibility that the instrument isolates exogenous variation in peer behavior. Conversely, if sorting occurs based on unobserved factors related to both peers' fathers' drinking and academic performance, the instrument may be endogenous. Overall, while the strategy is well-thought out and has potential.