FEM 11087 - Applied Microeconometrics

Assignment 1: Empirical Analysis

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

Group 23

Kees-Piet Barnhoorn
Tyler McGee
Andres Pinon
Jolien Schaeffers

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

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

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 ≥ 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
```

```
     Race: |
     Black |
   (1=Yes, |            overweight
     0=No) |         0          1          . |     Total
-----------+---------------------------------+----------
         0 |        87        296         11 |       394
           |     22.08      75.13       2.79 |    100.00
-----------+---------------------------------+----------
```

[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)

```
         1 |        13         50          2 |        65
           |     20.00      76.92       3.08 |    100.00
-----------+---------------------------------+----------
         . |         1          0          0 |         1
           |    100.00       0.00       0.00 |    100.00
-----------+---------------------------------+----------
     Total |       101        346         13 |       460
           |     21.96      75.22       2.83 |    100.00
```

```
     Race: |
     Black |
   (1=Yes, |              obese
     0=No) |         0          1          . |     Total
-----------+---------------------------------+----------
         0 |       225        158         11 |       394
           |     57.11      40.10       2.79 |    100.00
-----------+---------------------------------+----------
         1 |        28         35          2 |        65
           |     43.08      53.85       3.08 |    100.00
-----------+---------------------------------+----------
         . |         1          0          0 |         1
           |    100.00       0.00       0.00 |    100.00
-----------+---------------------------------+----------
     Total |       254        193         13 |       460
           |     55.22      41.96       2.83 |    100.00
```

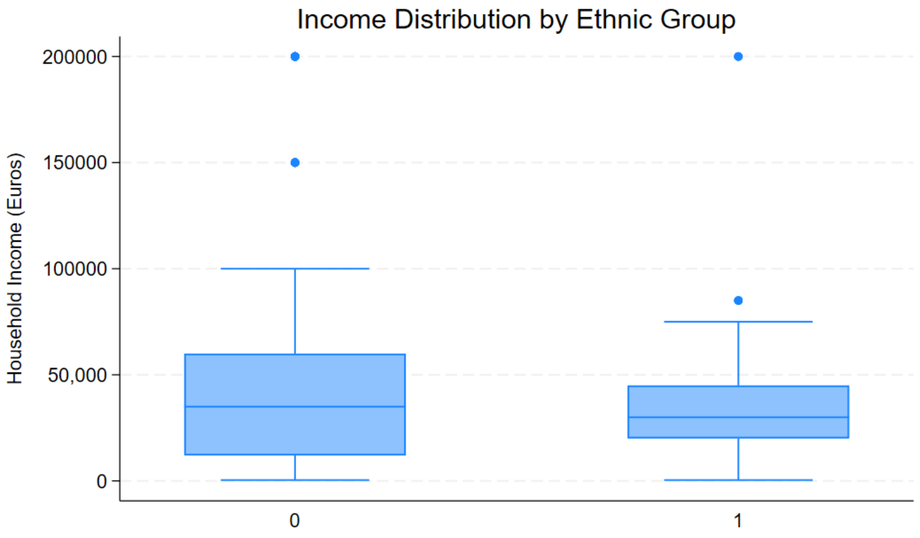
As we can see from the output, the rate of overweight is similar amongst ethnic groups (75.1% vs 76.9%); however, obesity rates are markedly higher in black individuals (53.8% vs 40.1%). This suggests that, while the total overweight rates are similar, black overweight individuals are more likely to 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)
```

Looking at , 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.*

Two main concerns could arise due to the missing BMI values in our sample. Firsly, missing BMI values effectively reduce the sample size. Secondly, if the BMI missingness is not random, but is instead related to other variables in our model, it could bias the regression estimates.

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

```
     Race: |
     Black |
   (1=Yes, |       bmi_miss
     0=No) |         0          1 |     Total
-----------+----------------------+----------
         0 |       383         11 |       394 
           |     97.21       2.79 |    100.00 
-----------+----------------------+----------
         1 |        63          2 |        65 
           |     96.92       3.08 |    100.00 
-----------+----------------------+----------
         . |         1          0 |         1 
           |    100.00       0.00 |    100.00 
-----------+----------------------+----------
     Total |       447         13 |       460 
           |     97.17       2.83 |    100.00 
```

The above example shows how BMI missingness would not be related to the `black` variable in our sample.

*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 | income > 500000
list height weight if height < 100 | height > 250
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 | income > 500000)
gen flag_height = (height < 100 | height > 250)
gen flag_bmi = (bmi < 10 | bmi > 60)
drop if flag_income == 1 | flag_height == 1 | flag_bmi == 1
```

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:

$$income = \beta_0 + \beta_1 BMI + \beta_2 black + u \tag{1}$$

Using **`reg`** with robust SE estimators:

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

```
Linear regression                               Number of obs     =        445
                                                F(2, 442)         =       1.07
                                                Prob > F          =     0.3450
                                                R-squared         =     0.0047
                                                Root MSE          =      39216

------------------------------------------------------------------------------
```

```
             |               Robust
      income | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         bmi |  -307.3532   277.6654    -1.11   0.269    -853.0618    238.3554
       black |  -4839.756   5108.111    -0.95   0.344    -14878.96    5199.447
       _cons |   51459.45   8992.888     5.72   0.000     33785.32    69133.59
------------------------------------------------------------------------------
```

*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 yearly income would decrease by 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 again a negative relationship with the dependent variable. The magnitude of the coefficient, 4839.756, implies a decrease of roughly 4840$ in annual income if the 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 it contains 0, we cannot reject the null hypothesis that BMI has no effect on income at the 5% significance level. This aligns with the conclusions we derived from 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, ceteris paribus. We can calculate it as:

$$\text{Relative Magnitude} = \text{Coefficient}/\text{Mean Income} \tag{2}$$

$$-4839.75/41,514.35 = -0.1165 \tag{3}$$

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 ***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
```

```
Linear regression                               Number of obs     =        445
                                                F(3, 441)         =       0.99
                                                Prob > F          =     0.3967
                                                R-squared         =     0.0049
                                                Root MSE          =      39257

------------------------------------------------------------------------------
             |               Robust
      income | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         bmi |     390.96   2387.682     0.16   0.870    -4301.689    5083.609
```

```
        bmi2 |  -10.84818   35.06745    -0.31   0.757    -79.76827    58.07192
       black |  -4812.927   5110.758    -0.94   0.347     -14857.4    5231.542
       _cons |    40704.1   39289.77     1.04   0.301    -36514.35    117922.5
------------------------------------------------------------------------------
```

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

$$\text{Income} = 40,704 + 390.96 \times \text{BMI} - 10.85 \times \text{BMI}^2 \tag{4}$$

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

$$40,704 + 390.96 \times 20 - 10.85 \times 20 = 44,183.2 \tag{5}$$

while, in the second case, we get

$$40,704 + 390.96 \times 30 - 10.85 \times 30 = 42,667.8 \tag{6}$$

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?,

At BMI 15, the marginal effect would be positive but at BMI 35, the marginal effect becomes negative. Since a linear model cannot take this pattern into account, adding BMI^2 allows the slope of our model to change at different BMI values.

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

The specification that includes BMI^2 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 two.

Question 4 [0.7 points]

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

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

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

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

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).*

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?*

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

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

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.*

Question 7 [0.6 points]

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

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

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.

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.*

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

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]*

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

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.

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

Question 10 [0.3 points]

What is your preferred model? Why?

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?*