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] **who_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?*

We first create binary indicators for overweight and obesity status:

```
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 more clearly compare the prevalence of both categories by ethnic group.

The prevalence rates are computed using:

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

	overweight		
Race: Black (1=Yes, 0=No)	0	1	Total
No.%No.%No.%			
08722.7%29677.3%383100.0%			
11320.6%5079.4%63100.0%			
Total10022.4%34677.6%446100.0%			

[2] **gould_stata**

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

```
hist income, by(black) ///
    title("Income Distribution by Ethnic Group") ///
    xtitle("Household Income (Euros)") ///
    percent
```

```
graph box income, over(black) ///
        title("Income Distribution by Ethnic Group") ///
        ytitle("Household Income (Euros)")
```

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

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

*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.*

```
sum income
```

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:

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

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

*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**?*

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

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.

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.*

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

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

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?*