

### Домашнее задание №1.

Домашнее задание сдается в электронном виде в SmartLMS в виде двух файлов:  
pdf-файл и код.

В pdf-файле должен содержаться только аккуратно оформленный отчет без кода. При  
отсутствии письменного отчета всё ДЗ оценивается на **0 баллов**.

В каждом пункте, где осуществляется проверка гипотез, построение доверительных  
интервалов и т.д., должны быть прописаны гипотезы, статистики и результаты расчетов  
(даже если вы оцениваете их с помощью готовых функций).

При обнаружении списывания работа обнуляется.

Дедлайн: 23:59, 16 марта 2025 года.

1. **(15 баллов)** В данной задаче мы будем исследовать взаимосвязь между весом новорожденных и потреблением сигарет матерями во время беременности (Mullahy, J. (1997), Instrumental Variable Estimation of Count Data Models: Application to Models of Cigarette Smoking Behavior, Review of Economics and Statistics 79, 586-593).

В файле *bwght.dta* представлены следующие данные:

- faminc: 1988 family income, \$1000s
- cigtax: cig. tax in home state, 1988
- cigprice: cig. price in home state, 1988
- bwght: birth weight, ounces
- fatheduc: father's yrs of educ
- motheduc: mother's yrs of educ
- parity: birth order of child
- male: =1 if male child
- white: =1 if white
- cigs: cigs smoked per day while preg
- lbwght: log of bwght
- bwghtlbs: birth weight, pounds
- packs: packs smoked per day while preg
- lfaminc: log(faminc)

Вы планируете оценить следующую модель:

$$lbwght_i = \beta_0 + \beta_1 \cdot male_i + \beta_2 \cdot parity_i + \beta_3 \cdot lfaminc_i + \beta_4 \cdot cigs_i + \varepsilon_i.$$

- (a) **(2 балл)** В чем может быть проблема при использовании *OLS* для оценки коэффициентов приведенной выше модели?

- (b) **(2 балл)** Предположим, у вас есть данные о средней цене сигарет в стране проживания. Поможет ли эта информация определить истинные параметры модели? Порассуждайте об этом.
- (c) **(2 балла)** Оцените с помощью OLS модель из пункта (a). Проинтерпретируйте полученные результаты.
- (d) **(3 балла)** Используя в качестве инструментальной переменной среднюю стоимость сигарет (*cigprice*), оцените модель из пункта (a) с помощью 2SLS. Сравните полученный результат с результатом из пункта (c).
- (e) **(4 балла)** Каким свойствам должна удовлетворять инструментальная переменная? На уровне значимости 5% проверьте их для инструментальной переменной из пункта (d), описав подробно используемые тесты.
- (f) **(2 балла)** На основе результатов пункта (e) дайте рекомендацию по оцениванию влияния курения матери в течение беременности на вес новорожденных.
2. **(30 баллов)** В рамках данной задачи вы хотите оценить эффект воздействия на будущую заработную плату ( $Y$ ) от посещения курса по эконометрике ( $D$ ) в университете. Определим потенциальные исходы следующим образом:

$$Y_i(0) = \alpha + \varepsilon_{i0},$$

$$Y_i(1) = \alpha + z_i + \varepsilon_{i1},$$

где  $E(\varepsilon_{i1}) = E(\varepsilon_{i0}) = 0$ , а  $D_i$  является индикатором воздействия:

$$D_i = \begin{cases} 1, & \text{если } i \text{ посещал курс по эконометрике} \\ 0, & \text{если } i \text{ не посещал курс по эконометрике,} \end{cases}$$

а  $z_i$  определяется следующим образом:

$$z_i = \begin{cases} 0, & \text{если } Z_i = 0 \\ z > 0, & \text{если } Z_i = 1 \end{cases}$$

где  $z$  — некоторая известная константа, а  $Z_i$  — это дамми на прохождение предварительного курса по линейной алгебре.

В выборке содержится информация о всех индивидах, посещавших и не посещавших курс по эконометрике, размеры групп которых обозначены как  $n_1$  и  $n_0$  соответственно. Однако наблюдаемыми наблюдениями являются  $D_i$  и

$$Y_i = Y_i(0) + D_i(Y_i(1) - Y_i(0)).$$

- (a) **(6 баллов)** Вычислите средний эффект воздействия (ATE – Averaged Treatment

Effect), средний эффект воздействия на подвергнутых воздействию (ATET – Averaged Treatment Effect on the Treated) и средний эффект воздействия на неподвергнутых воздействию (ATEU – Averaged Treatment Effect on the Untreated).

- (b) **(6 баллов)** При каких условиях три эффекта воздействия (ATE, ATET и ATEU) совпадают?
- (c) **(6 баллов)** При выполнении условия, что  $E(\varepsilon_{i0}|D_i) = 0$  какие из средних эффектов (ATE, ATET, ATEU) можно оценить? Опишите, как бы вы использовали имеющиеся данные для их оценки.
- (d) **(6 баллов)** Пусть теперь переменная воздействия определяется следующим образом:

$$D_i = \begin{cases} 1, & \text{если } z_i > 0 \\ 0, & \text{если } z_i = 0 \end{cases}$$

При сохранении предположения из пункта (c), какие из средних эффектов (ATE, ATET, ATEU) можно оценить? Опишите, как можно использовать данные для их оценки.

- (e) **(6 баллов)** Теперь предположим, что участие определяется следующим образом:

$$D_i = \begin{cases} 1, & \text{если } z_i - \theta_i \geq 0 \\ 0, & \text{если } z_i - \theta_i < 0 \end{cases}$$

где  $\theta_i \perp \varepsilon_{i1}, \varepsilon_{i0}$  и  $\theta \sim U(0, 2z)$ . Вычислите ATE, ATET и ATEU. Какие из этих эффектов можно оценить при выполнении условия  $E(\varepsilon_{i0}|D_i) = 0$ ? Опишите, как можно использовать данные для их оценки.

3. **(15 баллов)** Рассмотрим следующую систему одновременных уравнений:

$$y_{1i} = \alpha y_{2i} + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \varepsilon_{1i}, \quad (1)$$

$$y_{2i} = \beta y_{1i} + \gamma_3 z_{3i} + \varepsilon_{2i}. \quad (2)$$

- (a) **(3 балла)** Решите приведенную форму для  $y_{1i}$  и  $y_{2i}$ .
- (b) **(3 балла)** Какие регрессоры являются эндогенными? Обоснуйте свой ответ.
- (c) **(5 баллов)** Какие структурные уравнения идентифицируемы? Обоснуйте свой ответ.
- (d) **(4 балла)** Объясните, какую переменную можно использовать в качестве инструмента для  $y_{1i}$  в уравнении (2)? Каким условиям должна удовлетворять приведенная форма для  $y_{1i}$ , чтобы существовал валидный инструмент?

4. (15 баллов) Даны  $M$  регрессионных уравнений:

$$y_1 = X_1\beta_1 + u_1,$$

$$y_2 = X_2\beta_2 + u_2,$$

.....,

$$y_M = X_M\beta_M + u_M,$$

где  $y_i$  —  $n \times 1$  вектор зависимых переменных,  $X_i$  —  $n \times k_i$  — матрица независимых переменных,  $\beta_i$  —  $k_i \times 1$  — вектор неизвестных параметров,  $u_i$  —  $n \times 1$  — вектор случайных ошибок,  $i = 1, \dots, M$ . Будем предполагать, что  $\mathbb{E}(u_i) = 0$ ,  $\mathbb{E}(u_{is}u_{jt}) = \sigma_{ij}$  при  $s = t$  и 0 в противном случае.

**Теорема (Эквивалентность оценок доступного ОМНК (FGLS) и МНК (OLS) для систем внешне несвязанных уравнений SUR):** Если  $X_1 = X_2 = \dots = X_M$ , то есть, если во всех уравнениях SUR используется один и тот же набор регрессоров, значит, оценки OLS и GLS совпадают.

Приведите доказательство данной теоремы.

*Подсказка: используйте матричную запись и произведение Кронекера.*

5. (25 баллов) Рассмотрим набор данных *class.dta*, содержащий информацию о размере классов (*classsize*), средних результатах тестов по математике (*avgmath*) и тестов по вербальным навыкам (*avgverb*) для 2019 учащихся пятых классов в 1002 государственных школах Израиля, а также данные о численности учащихся в параллели (*enrollment*) в соответствующей школе и проценте малообеспеченных учеников (*disadv*).

В Израиле действует правило, согласно которому размер класса не может превышать 40 учеников. Когда численность учащихся достигает 41, школа должна открыть второй класс, затем третий класс при 81 учениках и так далее. Это вызывает резкие скачки в размерах классов при кратных 40 значениях.

Ограничьте выборку школами с численностью учащихся (*enrollment*) от 20 до 60. Создайте дамми переменную *large\_cohort*, указывающую на первую границу разрыва при 40 учениках.

- (a) (3 балла) Оцените влияние размера класса (*classsize*) на результаты тестов по математике (*avgmath*), используя МНК без каких-либо контролирующих переменных, а затем добавьте процент малообеспеченных учеников (*disadv*) в классе и общее количество учащихся (*enrollment*) в качестве контрольных переменных. Проинтерпретируйте полученные результаты.
- (b) (3 балла) Используйте МНК для оценки влияния обучения в большом классе (*largecohort*) на результаты тестов по математике с помощью четкого разрывно-

го дизайна (Sharp RDD) с учетом размера класса. Добавьте также контрольные переменные: процент малообеспеченных учеников и линейный тренд численности учащихся в параллели.

- (c) **(3 балла)** Представьте графически результаты пункта (b). Проинтерпретируйте.
- (d) **(3 балла)** Если метод разрывного дизайна является корректным, коэффициент интересующего нас параметра не должен значительно изменяться при включении или исключении контрольных переменных. Проверьте это утверждение.
- (e) **(3 балла)** Изобразите результаты оценивания из пункта (d) и дайте интерпретацию.
- (f) **(3 балла)** Проверьте устойчивость полученных результатов, заменив в модели из пункта (b) линейный тренд численности учащихся на квадратичный.
- (g) **(3 балла)** Проверьте устойчивость полученных результатов с помощью плацебо-теста: выполните анализ разрыва в регрессии (Sharp RDD), используя в качестве зависимой переменной процент малообеспеченных учеников. Результат представьте в виде оценки регрессии и графически.
- (h) **(4 балла)** Обратите внимание, что не все школы действовали согласно правилам при создании классов. Предложите метод оценки эффекта от обучения в большом классе на результаты теста по математике, аргументировав выбор. Запишите спецификацию модели, которую Вы бы оценили данным методом. По возможности оцените её.