Домашнее задание №1.

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 smked per day while preg
   - lbwght: log of bwght
   - bwghtlbs: birth weight, pounds
   - packs: packs smked 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?

Решение:

Очевидно, что самая большая проблема, которая может быть для OLS это эндогенность. Тут она может присутствовать из-за многих причин, например элементарно ошибка в измерении регрессоров, то есть наших факторов: в нашей модели присутствует регрессор $cigs$, который спокойно может не соответствовать действительности из-за неточности ответов респондентов, доход семьи тоже можно как и приукрасить, так и не договорить, точное значение мы вряд ли получим.

Также причиной может быть и пропущенная переменная, которая коррелированна с включенными в модель регрессорами, например наличие каких-либо менитальных проблем со здоровьем или иные факторы (стресс, который в том числе может влиять на вес ребенка и курение могут зависить от огромного количества неизмеренных факторов).

Проблемы с OLS в таком случае будут страшными: наши оценки будут смещены и несостоятельны, а такое эконометрика не прощает.

(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.

Решение:

Факторы нашей модели: пол, порядок рождения, логарифм семейного дохода и количество выкуренных сигарет в день во время беременности.

По первым двум все понятно, никакой информации, что касается семейного дохода, то тут очевидна возможность наличия корреляции со средней ценой сигарет в стране проживания: в странах с более высоким уровнем доходов уровень цен товаров и услуг тоже более высокий. Также можно в теории предположить обратную зависимость между уровнем цен на сигареты (возможное введение повышенных акцизов на табачную продукцию, как принимаемую меру по борьбе с курением) и количеством выкуренных сигарет, но проблема может быть в том, что уровень цен сигарет высокий из-за высокого уровня жизни и цен в целом, что не дает никакой гарантии на обратную связь и может спокойно означать, что доход у семьи большой и они могут курить сколько захотят.

В общем только на основании среднего уровня цен без анализа других факторов не стоит делать никаких выводов по поводу возможности помочь определить параметры модели, можно лишь понадеяться на наличие какой никакой, но корреляции.

(c) **(2 балла)** Оцените с помощью OLS модель из пункта (a). Проинтерпретируйте полученные результаты.

Решение:

OLS Regression Results

Dep. Variable:	lbwght	R-squared:	0.035
Model:	OLS	Adj. R-squared:	0.032
Method:	Least Squares	F-statistic:	12.55
Date:	Sun, 16 Mar 2025	Prob (F-statistic):	4.90e-10
Time:	17:17:32	Log-Likelihood:	356.03
No. Observations:	1388	AIC:	-702.1
Df Residuals:	1383	BIC:	-675.9
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.6756	0.022	213.681	0.000	4.633	4.719
male	0.0262	0.010	2.601	0.009	0.006	0.046
parity	0.0147	0.006	2.600	0.009	0.004	0.026
lfaminc	0.0180	0.006	3.233	0.001	0.007	0.029
cigs	-0.0042	0.001	-4.890	0.000	-0.006	-0.003

Omnibus:	614.841	Durbin-Watson:	1.931
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6025.606
Skew:	-1.799	Prob(JB):	0.00
Kurtosis:	12.552	Cond. No.	29.2

Рис. 1. Результаты модели

Как видно на классическом уровне значимости в 5% все коэффициенты оказались значимы.

Уравнение модели:

$lbwght = 4.676 + 0.026 * male + 0.015 * parity + 0.018 * lfaminc + -0.004 * cigs$

Интерпретация:

Мальчики в среднем весят на $(e^{0.026} - 1) \cdot 100\% \approx 2.6\%$, больше, чем девочки. Также каждый ребенок в семье в среднем весит на 1.5% больше предыдущего. При увеличении количества выкуренных сигарет в день во время беременности на единицу вес ребенка снижается в среднем на $\left((e^{0.004} - 1) \cdot 100\% \approx -0.39\right)$ 0.39%. Также при увеличении семейного дохода на 1% вес ребенка в среднем увеличивается на 0.018%.

(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).

Решение:

Эээээ нам нужно выбрать эндогенную переменную, для которой будем использовать инструмент. Выбираем cigs и применяем к ней инструмент cigprice. Результаты следующие:

```
                          IV-2SLS Estimation Summary
==============================================================================
Dep. Variable:                  lbwght   R-squared:                     -1.8118
Estimator:                     IV-2SLS   Adj. R-squared:                -1.8199
No. Observations:                 1388   F-statistic:                    10.018
Date:                Sun, Mar 16 2025   P-value (F-stat)                0.0401
Time:                        18:03:44   Distribution:                  chi2(4)
Cov. Estimator:                robust

                             Parameter Estimates
==============================================================================
            Parameter  Std. Err.     T-stat    P-value    Lower CI    Upper CI
------------------------------------------------------------------------------
const          4.4679     0.2559     17.463     0.0000      3.9664      4.9693
male           0.0298     0.0172     1.7348     0.0828     -0.0039      0.0635
parity        -0.0012     0.0253    -0.0489     0.9610     -0.0509      0.0484
lfaminc        0.0636     0.0570     1.1172     0.2639     -0.0480      0.1753
cigs           0.0399     0.0556     0.7173     0.4732     -0.0690      0.1488
==============================================================================

Endogenous: cigs
Instruments: cigprice
Robust Covariance (Heteroskedastic)
Debiased: False
```

Рис. 2. Результаты модели с инструментом cigprice

Ну что-то совсем странное, R^2 у нас улетел в отрицательное значение. Регрессоры стали статистически незначимыми. Эффект от cigs стал положительным, то есть по нему вес ребенка будет расти, что вообще неадекватно и противоречит OLS модели. В итоге качество упало, все незначимое и результаты неадекватны.

(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.

Решение:

Инструментальные переменные должны быть валидными (то есть экзогенными) и релевантными.

Про валидность мы ничего не можем сделать, Сарган нам недоступен (так как инструментальная переменная всего одна), а Хаусман не даст никакой точной информации. Все равно сделаем его, потому что больше ничего не можем.

$$H_0 : plim\frac{X'\varepsilon}{n} = 0$$

$$H_1 : plim\frac{X'\varepsilon}{n} \neq 0$$

Тестовая статистика:

$(\hat{\beta}_{IV} - \hat{\beta}_{\text{МНК}})'(V(\hat{\beta}_{IV}) - V(\hat{\beta}_{\text{МНК}}))^{-1}(\hat{\beta}_{IV} - \hat{\beta}_{\text{МНК}}) \sim \chi^2_k$

Получаем результат статистики = 1.9186, p-value = 0.1662 > 0.05 для 5% уровня значимости, то есть H_0 не отвергается, возможно вообще все было нормально.

Для релевантности проведем тест F-статистика первого шага.

То есть оцениваем нашу эндогенную переменную экзогенные и инструмент. Получаем:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   cigs   R-squared:                       0.030
Model:                            OLS   Adj. R-squared:                  0.028
Method:                 Least Squares   F-statistic:                     10.86
Date:                Sun, 16 Mar 2025   Prob (F-statistic):           1.14e-08
Time:                        18:55:09   Log-Likelihood:                -4428.2
No. Observations:                1388   AIC:                             8866.
Df Residuals:                    1383   BIC:                             8892.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.7482      2.080      1.321      0.187      -1.332       6.828
male          -0.0945      0.317     -0.298      0.766      -0.717       0.527
parity         0.3630      0.178      2.044      0.041       0.015       0.711
lfaminc       -1.0527      0.174     -6.051      0.000      -1.394      -0.711
cigprice       0.0155      0.016      1.001      0.317      -0.015       0.046
==============================================================================
Omnibus:                     1025.554   Durbin-Watson:                   1.945
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            14470.841
Skew:                           3.423   Prob(JB):                         0.00
Kurtosis:                      17.260   Cond. No.                     1.72e+03
==============================================================================
...
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.72e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
<F test: F=1.0018004472161448, p=0.31705033061915966, df_denom=1.38e+03, df_num=1>
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Рис. 3. F-статистика первого шага

То есть F-статистика = 1.001, p-value = 0.317. Получаем что F < 10, p-value > 0.05 и на уровне значимости 5% мы отвергаем гипотезу о значимости cigprice, то есть инструмент не является релевантным.

(f) **(2 балла)** На основе результатов пункта (e) дайте рекомендацию по оцениванию влияния курения матери в течение беременности на вес новорожденных.

Решение:

Исходя из полученных нами результатов, cigprice не является хорошим инструментом. Возможно стоит взять cigtax, так как он может более хорошо быть связанным с cigs, как раз повышение акцизов может свидетельствовать об ужесточении политики к табачному производству и на борьбу с курением. Также стоило собирать данные по большему числу факторов, которые могли бы отражать cigs. Например дамми-переменные "есть ли курящие родственники"или "наличие ОКР/СДВГ/невротик ли".

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:

$$y_{1i} = \alpha y_{2i} + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \varepsilon_i, \quad (1)$$
$$y_{2i} = \beta y_{1i} + \gamma_3 z_{3i} + \varepsilon_i. \quad (2)$$

(a) **(3 балла)** Решите приведенную форму для y_{1i} и y_{2i}.

Решение:

Перепишем систему в матричном виде:

$$y_i = \begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix}, z_i = \begin{bmatrix} z_{1i} \\ z_{2i} \\ z_{3i} \end{bmatrix}, \varepsilon_i = \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{bmatrix},$$
$$B = \begin{bmatrix} 1 & -\alpha \\ -\beta & 1 \end{bmatrix}, \Gamma = \begin{bmatrix} -\gamma_1 & -\gamma_2 & 0 \\ 0 & 0 & -\gamma_3 \end{bmatrix}$$

Тогда система будет: $By_i + \Gamma x_i = \varepsilon_i$

Приведенный вид: $y_i = -B^{-1}\Gamma x_t + B^{-1}\varepsilon_i$

$$y_i = \frac{1}{\alpha\beta - 1} \begin{bmatrix} 1 & \alpha \\ \beta & 1 \end{bmatrix} \begin{bmatrix} -\gamma_1 & -\gamma_2 & 0 \\ 0 & 0 & -\gamma_3 \end{bmatrix} \begin{bmatrix} z_{1i} \\ z_{2i} \\ z_{3i} \end{bmatrix} + \frac{1}{1-\alpha\beta} \begin{bmatrix} 1 & \alpha \\ \beta & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{bmatrix}$$

$$\begin{bmatrix} y_{1i} \\ y_{i2} \end{bmatrix} = \frac{1}{\alpha\beta - 1} \begin{bmatrix} -\gamma_1 & -\gamma_2 & -\alpha\gamma_3 \\ -\beta\gamma_1 & -\beta\gamma_2 & -\gamma_3 \end{bmatrix} \begin{bmatrix} z_{1i} \\ z_{2i} \\ z_{3i} \end{bmatrix} + \frac{1}{1-\alpha\beta} \begin{bmatrix} \varepsilon_{1i} + \alpha\varepsilon_{2i} \\ \beta\varepsilon_{1i} + \varepsilon_{2i} \end{bmatrix}$$

$y_{1i} = \frac{z_{1i}\gamma_1 + z_{2i}\gamma_2 + \alpha\gamma_3 z_{3i} + \varepsilon_{1i} + \alpha\varepsilon_{2i}}{1-\alpha\beta}$

$y_{2i} = \frac{\beta z_{1i}\gamma_1 + \beta z_{2i}\gamma_2 + \gamma_3 z_{3i} + \beta\varepsilon_{1i} + \varepsilon_{2i}}{1-\alpha\beta}$

(b) **(3 балла)** Какие регрессоры являются эндогенными? Обоснуйте свой ответ.

Решение:

Экзогенные: z_i так как они входят в каждое уравнение в уникальной комбинации и не подразумевают двустороннюю зависимость, как y_{1i} и y_{i2}

Эндогенные же, то есть входящие в каждое уравнение и таким образом зависящие от ошибки: y_{1i} и y_{i2}

(c) **(5 баллов)** Какие структурные уравнения идентифицируемы? Обоснуйте свой ответ.

Решение:

Давайте проверим порядковое и ранговые условия:

Порядковое условие $r_i \geq m - 1$, то есть $r_i \geq 1$?

(r_i – кол-во ЛНЗ огр-й, m – кол-во уравнений (эндогенных регрессоров))

Посчитаем кол-во ЛНЗ ограничений.

Для уравнения (1) огр-я: $\gamma_3 = 0$

Для уравнения (2) огр-я: $\gamma_1 = 0, \gamma_2 = 0$

Таким образом $r_2 = 2 > r_1 = 1 \geq m - 2 = 1$, то есть порядковое условие выполнено для обоих уравнений.

Ранговое условие: составим табличку и воспользуемся лайфхаком, чтобы не мучаться с Ашками и Фишакми. В этой таблице записаны коэффициенты всех регрессоров первого и второго уравнений, нули это и будут наши ограничения.

y_1	y_2	z_1	z_2	z_3
1	$-\alpha$	$-\gamma_1$	$-\gamma_2$	0
$-\beta$	1	0	0	$-\gamma_3$

Для первого уравнения берем все стобцы где есть нули в первой строке, это последнй столбец:

z_3
0
$-\gamma_3$

Вычеркиваем 0 и получаем просто матрицу 1×1 $rk(-\gamma_3) = 1$, так как все исключающие ограничения мы учли. Таким образом первое уравнение точно идентифицируемо.

Аналогично берем все столбцы где есть нули для второй строки (для второго уравнения).

Получим

z_1	z_2
$-\gamma_1$	$-\gamma_2$
0	0

Получаем строку: $rk\left(-\gamma_1\ -\gamma_2\right) = 1$. Ранговое условие для второго уравнения тоже выполнено, оно сверхидентифицируемо так как $r_2 > m - 1 = 1$

Таким образом получаем, что оба уравнения идентифицируемы: первое точно так как поряковое условие выполняется со знаком равенства, второе сверхиндетифицируемо, так как порядковое условие выполняется со строгим знаком.

(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?

Решение:

Для существования валидного инструмента необходимо, чтобы уравнение было идентифицируемо, также в нашем случае $1 - \alpha\beta \neq 0$, так как иначе решений не найдется. В качестве валидного инструмента берем либо z_1 либо z_2, так как они не коррелируют с ошибкой ε_{2i} (экзогенны для y_{2i}), но тем не менее коррелируют с y_{1i} (при условии, что $\gamma_i \neq 0$ для i-го выбранного инструмента, $i \in \{1, 2\}$), то есть они валидные и релевантные.

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, β_i — $k_i \times 1$ — вектор неизвестных параметров, u_i — $n \times 1$ — вектор случайных ошибок, $i = 1, ..., M$. 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 $X_1 = X_2 = ... = X_M$, 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.

Решение:

Так как $X_1 = X_2 = \cdots = X_n = X_0$, то $k_1 = \cdots = k_n = k$

Запишем всю систему в матричном виде:

$$y_{Mn\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_M \end{bmatrix}, X_{Mn\times Mk} = \begin{bmatrix} X_0 & 0 & 0 & 0 \\ 0 & X_0 & 0 & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & X_0 \end{bmatrix} = I_M \otimes X_{0_{n\times k}},$$

$$\beta_{Mk\times 1} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_M \end{bmatrix}, \varepsilon_{Mn\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_M \end{bmatrix}, \mathbb{E}(\varepsilon_i\varepsilon_j) = \sigma_{ij}I_n$$

Тогда система принимает стандартный вид: $y = X\beta + \varepsilon$

$$\Omega_{Mn\times Mn} = \mathbb{E}(\varepsilon\varepsilon^T) = \begin{bmatrix} \mathbb{E}(\varepsilon_1\varepsilon_1) & \mathbb{E}(\varepsilon_1\varepsilon_2) & \cdots & \mathbb{E}(\varepsilon_1\varepsilon_M) \\ \mathbb{E}(\varepsilon_2\varepsilon_1) & \mathbb{E}(\varepsilon_2\varepsilon_2) & \cdots & \mathbb{E}(\varepsilon_2\varepsilon_M) \\ \cdots & \cdots & \cdots & \cdots \\ \mathbb{E}(\varepsilon_M\varepsilon_1) & \mathbb{E}(\varepsilon_M\varepsilon_2) & \cdots & \mathbb{E}(\varepsilon_M\varepsilon_M) \end{bmatrix} = \Sigma \otimes I_n$$

$$\Sigma_{M\times M} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1M} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{M1} & \sigma_{M2} & \cdots & \sigma_{MM} \end{bmatrix}$$

Рассмотрим оценки OLS и GLS:

$\beta^{OLS} = (X^T X)^{-1} X^T y$

$\beta^{GLS} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y = ()$

Распишем (матрицу перед y, если они сойдутся, то оценки совпадают)

(Я зуб даю все размерности сходятся я перепроверил на айпаде)

$$0LS : (X^T X)^{-1} X^T = ((I_M \otimes X_0)^T (I_M \otimes X_0))^{-1} (I_M \otimes X_0)^T =$$

$$((I_M + X_0^T)(I_M \otimes X_o))^{-1} (I_M \otimes X_0^T) = (I_M \otimes (X_0^T X_0))^{-1} (I_M \otimes X_0^T) =$$

$$= (I_M \otimes (X_0^T X_0)^{-1})(I_M \otimes X_0^T) = I_M \otimes (X_0^T X_0)^{-1} X_0^T$$

$$GLS : (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} = ((I_M \otimes X_0^T)(\Sigma^{-1} \otimes I_n)(I_M \otimes X_0))^{-1} \cdot$$

$$(I_M \otimes X_0^T)(\Sigma^{-1} \otimes I_n) = ((\Sigma^{-1} \otimes X_0^T)(I_M \otimes X_0))^{-1} (I_M \otimes X_0^T)(\Sigma^{-1} \otimes I_n) =$$

$$= (\Sigma^{-1} \otimes X_0^T X_0)^{-1} (I_M \otimes X_0^T)(\Sigma^{-1} \otimes I_n) = (\Sigma \otimes (X_0^T X_0)^{-1})(I_M \otimes X_0^T)(\Sigma^{-1} \otimes I_n)$$

$$= (\Sigma \otimes (X_0^T X_0)^{-1})(\Sigma^{-1} \otimes X_0^T) = I_M \otimes (X_0^T X_0)^{-1} X_0^T$$

Совпало))), значит GLS оценка в таком случае равна OLS оценке, в частности это произошло потому что мы смогли расписать блочную матрицу X через кронекерово произведение единичной матрицы и X_0, которая в частном случае нашем - матрица факторов для всех уравнений.

$\mathbb{Q}.\mathbb{E}.\mathbb{D}.$

5. **(25 баллов)** Рассмотрим набор данных *class.dta*, содержащий информацию о размере классов ($classize$), средних результатах тестов по математике ($avgmath$) и тестов по вербальным навыкам ($avgberb$) 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 ученике 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 ($classize$) 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 разрывного 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.