Добин Илья
ДЗ-4

Задание 1

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

a)

Проблема при использовании OLS: ***cigs*** может быть эндогенна. Скорее всего существуют неучтенные факторы, которые влияют как на вес новорожденного, так и на курение матери. Поэтому оценки могут быть несостоятельны.

b)

Средняя цена сигарет в стране может быть полезной информацией только в том случае, когда она коррелирует с количеством выкуренных сигарет матерями. Но на мой взгляд увеличение цены не обязательно будет сильно снижать потребление сигарет, так как этот товар вызывает сильную зависимость и зависимый человек готов покупать товар даже если он будет переоценен.

Но если же корреляция достаточная, то эта переменная может решать проблему эндогенности ***cigs***, так как логично, что она не влияет на другие факторы, т. е. экзогенна и вследствие наличия корреляции - релевантна.

Хотя экзогенность все-таки не очевидна. Возможно цена на сигареты как-то связана с условиями жизни в стране, что может влиять на например здоровье матери.

c)

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:	20:12:16	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

Все коэффициенты стат. значимы на уровне 1%. Курение отрицательно влияет на вес ребенка. Мальчики в среднем весят больше девочек. Порядковый номер родов ***parity*** и семейный доход ***lfaminc*** положительно влияют на вес. Но R^2 всего 0.035.

d)

Результаты первого шага:

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:	20:31:00	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]
Intercept	2.7482	2.080	1.321	0.187	-1.332	6.828
cigprice	0.0155	0.016	1.001	0.317	-0.015	0.046
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

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

F-статистика = 10.86, значит инструмент не слабый. Также интересно, что цена сигарет ***cigprice*** положительно влияет на количество выкуренных сигарет ***cigs***, но этот эффект не стат. значим.

Результаты второго шага:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.4679	0.153	29.231	0.000	4.168	4.768
cigs_hat	0.0399	0.032	1.243	0.214	-0.023	0.103
male	0.0298	0.010	2.840	0.005	0.009	0.050
parity	-0.0012	0.013	-0.096	0.924	-0.027	0.024
lfaminc	0.0636	0.034	1.890	0.059	-0.002	0.130

Наблюдаем положительный эффект курения на вес ребенка (в 2SLS коэффициент 0.0399, а в OLS был -0.0042), но коэффициент не значим. Видимо все-таки инструмент оказался слабым, а оценки 2SLS ненадежными.

e)

Уже упоминал в пункте a) - инструментальная переменная должна обладать 1. **релевантностью** - инструмент должен быть хорошо коррелированным с эндогенным регрессором, и 2. **экзогенностью** - инструмент не должен коррелировать с ошибкой в основном уравнении регрессии.

Проверка релевантности:

t-статистика инструмента 1.001 с pvalue 0.317, следовательно на уровне значимости 5% мы не отвергаем гипотезу о том, что коэффициент при инструменте равен нулю. => инструмент слабый.

Проверка экзогенности:

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.5890	0.066	69.256	0.000	4.459	4.719
male	0.0257	0.010	2.541	0.011	0.006	0.045
parity	0.0148	0.006	2.606	0.009	0.004	0.026
lfaminc	0.0172	0.006	3.073	0.002	0.006	0.028
cigs	-0.0042	0.001	-4.928	0.000	-0.006	-0.003
cigprice	0.0007	0.000	1.385	0.166	-0.000	0.002

Включили в модель cigprice. pvalue коэффициента 0.0007 равна 0.166 => на уровне значимости 5% не можем отвергнуть гипотезу о практически отстутствии влияния cigprice на вес ребенка. => инструмент экзогенный.

f)

2SLS оценки ненадежны, так как инструмент нерелевантный. Следовательно надо применять методы, устойчивые к слабым инструментам, или искать сильные инструменты. Очевидно, можно увеличить выборку, тогда инструменты станут мощнее и улучшить основную регрессию, то есть добавить новых признаков, чтобы R^2 был повыше и модель объясняла вес ребенка лучше.

Задание 5

a)

Регрессия 1:

OLS Regression Results

Dep. Variable:	avgmath	R-squared:	0.024
Model:	OLS	Adj. R-squared:	0.022
Method:	Least Squares	F-statistic:	17.03
Date:	Sun, 16 Mar 2025	Prob (F-statistic):	4.13e-05
Time:	21:42:16	Log-Likelihood:	-2620.5
No. Observations:	699	AIC:	5245.
Df Residuals:	697	BIC:	5254.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	56.6124	1.982	28.557	0.000	52.720	60.505
classize	0.3141	0.076	4.127	0.000	0.165	0.464

Размер класса ***classize*** положительно влияет на успеваемость. R^2 очевидно низкий и равен 0.024, то есть модель слабо объясняет разброс результатов.

Регрессия 2:

OLS Regression Results

Dep. Variable:	avgmath	R-squared:	0.311
Model:	OLS	Adj. R-squared:	0.308
Method:	Least Squares	F-statistic:	104.7
Date:	Sun, 16 Mar 2025	Prob (F-statistic):	6.26e-56
Time:	21:42:53	Log-Likelihood:	-2498.6
No. Observations:	699	AIC:	5005.
Df Residuals:	695	BIC:	5023.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	66.9999	2.050	32.689	0.000	62.976	71.024
classize	0.1313	0.066	1.997	0.046	0.002	0.260
disadv	-0.3323	0.020	-16.863	0.000	-0.371	-0.294
enrollment	0.0234	0.029	0.821	0.412	-0.033	0.079

classize по прежнему положительно влияет на успеваемость, однако pvalue выросло до 0.046 и коэффициент уже будет не значим на уровне значимости 1%. ***enrollment*** также оказывает положительный эффект, однако коэффициент точно не значим. ***disadv*** отрицательно влияет на результаты, что логично. R^2 увеличился до 0.311.

b)

OLS Regression Results

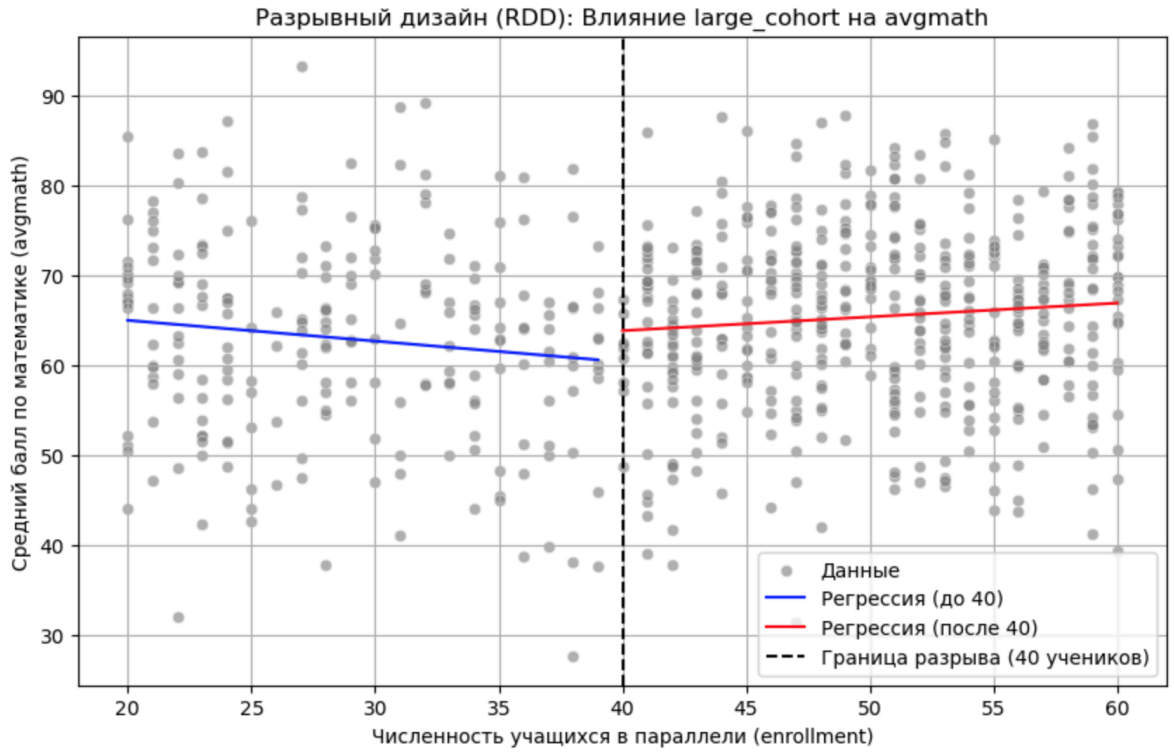
Dep. Variable:	avgmath	R-squared:	0.321
Model:	OLS	Adj. R-squared:	0.317
Method:	Least Squares	F-statistic:	82.09
Date:	Sun, 16 Mar 2025	Prob (F-statistic):	4.71e-57
Time:	21:59:07	Log-Likelihood:	-2493.6
No. Observations:	699	AIC:	4997.
Df Residuals:	694	BIC:	5020.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	68.8815	2.120	32.486	0.000	64.718	73.045
large_cohort	4.7644	1.497	3.183	0.002	1.826	7.703
classize	0.2192	0.071	3.090	0.002	0.080	0.358
disadv	-0.3319	0.020	-16.954	0.000	-0.370	-0.293
enrollment	-0.1467	0.060	-2.425	0.016	-0.265	-0.028

Коэффициент при large_cohort равен 4.764. Это значит, что при переходе через порог в 40 учеников средний балл по математике увеличивается на 5.69 баллов. Логично, так как при переходе размер классов становится меньше, следовательно учитель может уделить больше внимания каждому ученику. Коэффициент при classize остается положительным.

c)

Синяя линия - тренд для школ с численность учащихся до 40 и красная соответственно от 40. Слева у нас школы с одним классом, справа с двумя. Соответственно слева среднее число учеников в классе равно абсциссе, а справа абсциссе/2. Замечаем, что и слева, и справа: чем меньше среднее число учеников в классе, тем в среднем лучше результаты тестов.



d)

RDD без контрольных переменных:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	54.6491	2.072	26.377	0.000	50.581	58.717
large_cohort	2.5500	0.831	3.069	0.002	0.919	4.181
classize	0.3230	0.076	4.266	0.000	0.174	0.472

коэффициент large_cohort равен 2.55

RDD с enrollment:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	56.3565	2.362	23.861	0.000	51.719	60.994
large_cohort	4.9115	1.779	2.761	0.006	1.419	8.404
classize	0.3763	0.084	4.503	0.000	0.212	0.540
enrollment	-0.1078	0.072	-1.501	0.134	-0.249	0.033

коэффициент large_cohort равен 4.9115

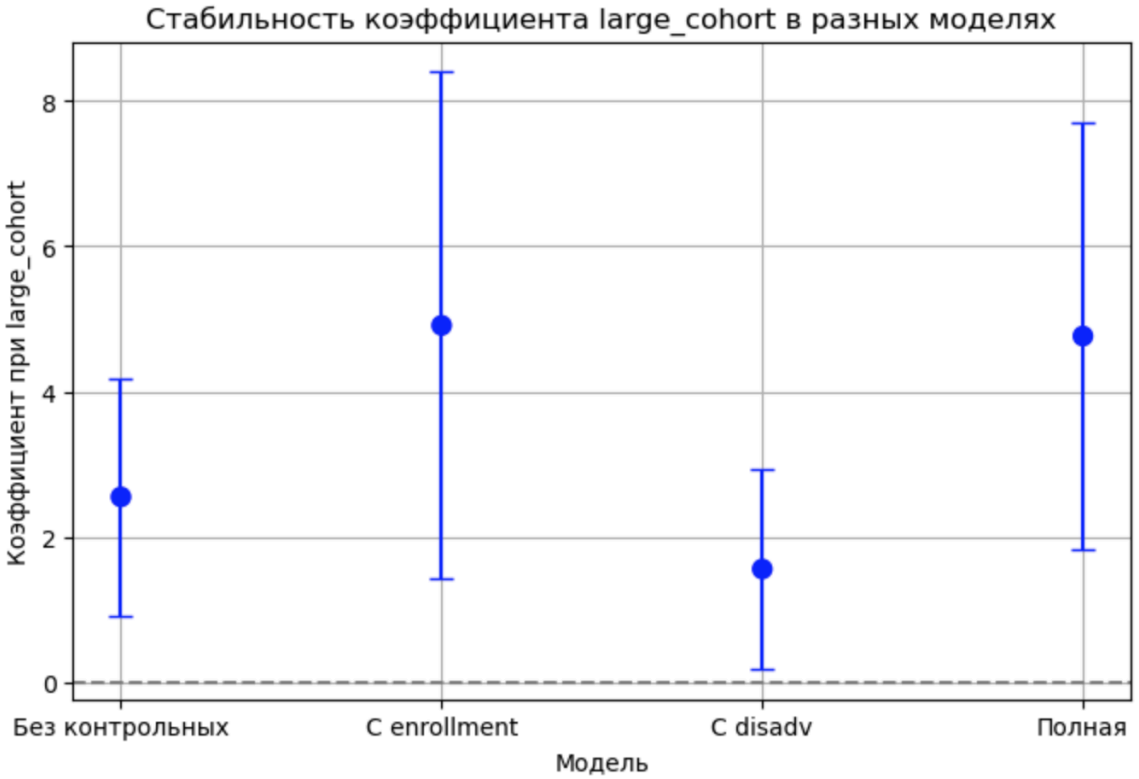
RDD с disadv:

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	66.4940	1.885	35.283	0.000	62.794	70.194
large_cohort	1.5569	0.703	2.213	0.027	0.176	2.938
classize	0.1477	0.065	2.281	0.023	0.021	0.275
disadv	-0.3301	0.020	-16.816	0.000	-0.369	-0.292

коэффициент large_cohort равен 1.5569

В полной модели же коэффициент large_cohort равен 4.7644
Видим, что коэффициент сильно меняется.

e)



Коэффициент везде значим, но меняется при добавлении контрольных переменных. Все доверительные интервалы > 0. То есть гарантированно наблюдается положительная связь между large_cohort и результатами тестов.

f)

OLS Regression Results

Dep. Variable:	avgmath	R-squared:	0.329
Model:	OLS	Adj. R-squared:	0.324
Method:	Least Squares	F-statistic:	67.95
Date:	Sun, 16 Mar 2025	Prob (F-statistic):	8.62e-58
Time:	22:45:04	Log-Likelihood:	-2489.5
No. Observations:	699	AIC:	4991.
Df Residuals:	693	BIC:	5018.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	80.7790	4.694	17.209	0.000	71.563	89.995
large_cohort	6.3771	1.594	4.001	0.000	3.247	9.507
classize	0.2541	0.072	3.547	0.000	0.113	0.395
disadv	-0.3297	0.019	-16.910	0.000	-0.368	-0.291
enrollment	-0.8249	0.246	-3.347	0.001	-1.309	-0.341
enrollment_sq	0.0077	0.003	2.837	0.005	0.002	0.013

Коэффициент при large_cohort увеличился до 6.37. Значит результаты устойчивы к изменению функциональной формы зависимости от enrollment. enrollment_sq статзначимый => реально нелинейная форма зависимости. При этом успеваемость сначала снижаются с ростом enrollment, а затем начинают расти.

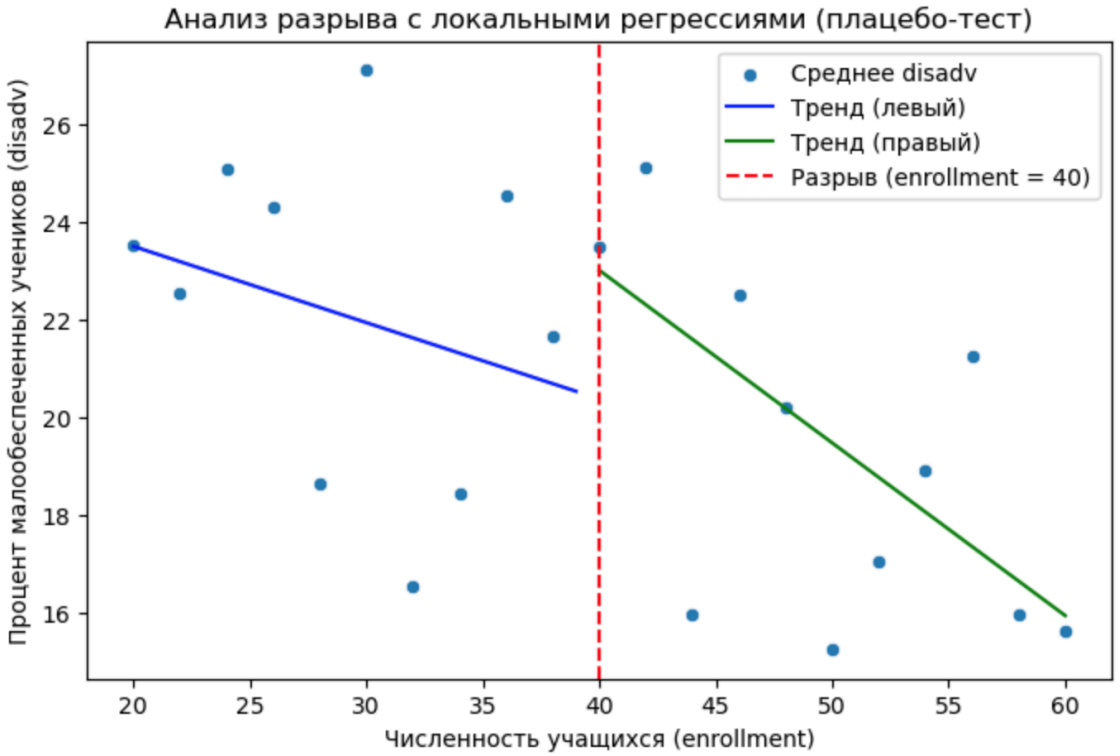
g)

OLS Regression Results

Dep. Variable:	disadv	R-squared:	0.033
Model:	OLS	Adj. R-squared:	0.029
Method:	Least Squares	F-statistic:	7.928
Date:	Sun, 16 Mar 2025	Prob (F-statistic):	3.33e-05
Time:	22:59:43	Log-Likelihood:	-2956.4
No. Observations:	699	AIC:	5921.
Df Residuals:	695	BIC:	5939.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	37.7362	3.851	9.799	0.000	30.175	45.297
large_cohort	-0.4432	2.900	-0.153	0.879	-6.137	5.251
classize	-0.4733	0.136	-3.474	0.001	-0.741	-0.206
enrollment	-0.1171	0.117	-1.000	0.318	-0.347	0.113

коэффициент large_cohort не стат. значим. Такой результат показывает, что наш RDD подход является валидным



№3.

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

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

a) $y_{1i} = \alpha(\beta y_{1i} + \gamma_3 z_{3i} + \varepsilon_{2i}) + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \varepsilon_{1i}$

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

$\Rightarrow y_{2i} = \beta\left(\alpha(\gamma_3 z_{3i} + \varepsilon_{2i} + \gamma_1 z_{1i} + \gamma_2 z_{2i} + \gamma_2 z_{2i} + \varepsilon_{1i}) \cdot \frac{1}{1-\alpha\beta}\right) + \gamma_3 z_{3i} + \varepsilon_{2i}$

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

	y_{1i}	y_{2i}	z_{1i}	z_{2i}	z_{3i}	~~Равносильное~~ Порядковое условие
(1)	1	$-\alpha$	$-\gamma_1$	$-\gamma_2$	0	$r_i \geq m-1$
(2)	$-\beta$	1	0	0	$-\gamma_3$	$2 > 1$ ✓

MUSIC

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

b) y_{1i}: эндогенный. Он коррелирован с ε_{2i}

y_{2i}: тоже эндогенный (корреляция с ε_{1i})

c) Порядковое условие уравнений выполнено,

⇒ надо еще проверить ранговое

(1): $rk(-\gamma_3) = 1$ — ранговое тоже выполнено

(2): $rk(-\gamma_1 - \gamma_2) = 2$

⇒ оба уравнения идентифицируемы

d) z_{1i} и z_{2i}, т.к. они не входят в (2); влияют на y_{1i} через приведённую форму.

Условия приведенной формы для Э валидного инструмента:

- нужна норм корреляция между инструментом и эндогенной y_{1i}:

⇒ $\gamma_1 \neq 0$ или $\gamma_2 \neq 0$

N4.

SUR: $\begin{cases} y_1 = X_1\beta_1 + \varepsilon_1 \\ \vdots \\ y_m = X_m\beta_m + \varepsilon_m \end{cases}$ $\quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$ $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix}$ $X = \begin{pmatrix} X_1 & & 0 \\ & \ddots & \\ 0 & & X_m \end{pmatrix}$

$E(\varepsilon_t | X_t) = 0$

$E(\varepsilon_{it}\varepsilon_{js}) = \begin{cases} \sigma_{ij}, & t = s \\ 0, & t \neq s \end{cases}$

$E(\varepsilon_i \varepsilon_j^T) = \sigma_{ij} I_n$

$E(\varepsilon\varepsilon^T) = \Omega = \Sigma \otimes I_n$, где $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots \\ \vdots & \ddots & \vdots \\ \cdots & & \sigma_{mm} \end{pmatrix}$

$\Omega^{-1} = \Sigma^{-1} \otimes I_n$

- GLS:

$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}y)$, где $\hat{\Omega} = \hat{\Sigma} \otimes I_n$

- OLS:

$e_1 = y_1 - X_1\hat{\beta}_1^{OLS}$

$\vdots$

$e_m = y_m - X_m\hat{\beta}_m^{OLS}$

$\hat{\sigma}_{ii} = \frac{RSS_i}{n} = \frac{e_i'e_i}{n}$

$\hat{\sigma}_{ij} = \widehat{cov}(\varepsilon_i, \varepsilon_j) = \frac{e_i'e_j}{n}$

DOBIN

Не то не от получилось

Th: $X_1 = \ldots = X_m \Rightarrow$ оценки OLS и GLS совпадают

□ $X_1 = \ldots = X_m \Rightarrow X = I_m \otimes X_0$

1) $X'\Omega^{-1}X = (I_m \otimes X_0)'(\Sigma^{-1} \otimes I_n)(I_m \otimes X_0) = (I_m' \otimes X_0')(\Sigma^{-1} \otimes I_n)(I_m \otimes X_0) = (I_m'\Sigma^{-1}I_n) \otimes (X_0'I_nX_0)$

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

$\Rightarrow (X'\Omega^{-1}X)^{-1} = \Sigma \otimes (X_0'X_0)^{-1}$

2) $X'\Omega^{-1}y = (I_m \otimes X_0)'(\Sigma^{-1} \otimes I_n)y = (I_m' \otimes X_0')(\Sigma^{-1} \otimes I_n)y$

$\Rightarrow \hat{\beta}_{GLS} = \left(\Sigma \otimes (X_0'X_0)^{-1}\right)\left((I_m' \otimes X_0')(\Sigma^{-1} \otimes I_n)y\right)$ (*)

$= ((I_m'\Sigma^{-1}) \otimes (X_0'I_n))y$

$= (\Sigma^{-1} \otimes X_0')y$

(*) $= \left(\Sigma \otimes (X_0'X_0)^{-1}\right)(\Sigma^{-1} \otimes X_0')y = ((\Sigma\Sigma^{-1}) \otimes ((X_0'X_0)^{-1}X_0'))y = (I_m \otimes (X_0'X_0)^{-1}X_0')y$

$\Rightarrow \hat{\beta}_{i,GLS} = (X_0'X_0)^{-1}X_0'y_i = \hat{\beta}_{i,OLS}$ $\forall i$ $\Rightarrow \hat{\beta}_{GLS} = \hat{\beta}_{OLS}$ ■

↑ МНК