Григораш Екатерина

Метод инструментальных переменных

Введение

В данной работе мы хотим проверить, что влияет на успеваемость студентов. В качестве зависимой переменной будет использоваться балл набранный на тесте (score), а среди независимых переменных находятся пол студента (gender), его/ее этничность (ethnicity), наличие высшего образования у отца и матери (fcollege и mcollege, соответственно), количество лет обучения (education) и доход семьи (income).

Перед тем как рассмотреть использование некоторых инструментальных переменных, проведем оценку модели в самом простом ее виде: с использованием зависимой переменной и всех независимых. Так как зависимая переменная является непрерывной, то мы будем использовать обычную линейную регрессию (МНК-регерссия). Результаты представлены в Таблице 1. Как мы видим, значения p-value всех переменных кроме дохода являются значимыми и оказывают эффект на успеваемость студента.

		Score	
Predictors	*Estimates*	*Std. Error*	*p*
(Intercept)	25.94	0.85	**<0.001**
education	1.96	0.06	**<0.001**
gender [female]	-1.14	0.21	**<0.001**
ethnicity [afam]	-6.55	0.29	**<0.001**
ethnicity [hispanic]	-4.29	0.28	**<0.001**
income [high]	0.01	0.25	0.957
fcollege [yes]	1.38	0.30	**<0.001**
mcollege [yes]	1.11	0.34	**<0.001**
Observations	4739		
R^2 / R^2 adjusted	0.328 / 0.327		

Таблица 1. Результаты линейной регрессии для оценки эффекта всех переменных на успеваемость.

Метод инструментальных переменных

Данный метод мы будем использовать, так как некоторые независимые переменные могут не оказывать прямого эффекта на успеваемость студента. Так, например, образование родителей (fcollege и mcollege) напрямую может не влиять на успеваемость студента, а влиять на количества лет, которые их ребенок отдает на образование. Объяснение может быть достаточно примитивным – если родители сами получали высшее образование и знают его ценность, то могут стимулировать их ребенка поступать таким же образом, отдавая обучению больше времени. Можно было бы еще в качестве инструментальной переменной использовать доход – однако она не значима (и сильно) в простой модели оценки эффекта, поэтому ее использовать мы не будем.

В данной работе мы реализуем метод инструментальных переменных с помощью двух способов: двухступенчатой модели (МНК-регрессия) и функции ivreg(). Последний позволит более точно рассчитать стандартные ошибки. Кроме того, чтобы сравнить модель, рассчитанную без инструментальных переменных, с моделями с инструментальными переменными, мы рассчитали эффект на успеваемость студента без учета образования родителей (см. Таблица 2). Именно эту модель мы и будем считать первой, с которой в дальнейшем будем сравнивать результаты метода инструментальных переменных.

		Score	
Predictors	*Estimates*	*Std. Error*	*p*
(Intercept)	24.85	0.84	**<0.001**
education	2.06	0.06	**<0.001**
gender [female]	-1.17	0.21	**<0.001**
ethnicity [afam]	-6.65	0.29	**<0.001**
ethnicity [hispanic]	-4.48	0.28	**<0.001**
income [high]	0.54	0.24	**0.025**
Observations	4739		
R^2 / R^2 adjusted	0.321 / 0.320		

Таблица 2. Результаты линейной регрессии для оценки эффекта всех переменных (кроме образования родителей) на успеваемость.

МНК-регрессия

Суть данного метода заключается в том, что мы проверяем корреляцию между нашей переменной education (кол-во лет на образование) и остальными независимыми переменными, среди которых есть наш инструмент (fcollege и mcollege). Если будет наблюдаться сильная корреляция между эндогенной переменной и инструментом, то значит, мы были правы, сказав, что прямого эффекта от образования родителей на успеваемость нет. А следовательно, используя образование родителей как инструмент, мы сможем более точно рассчитать истинный эффект от остальных переменных, в частности от количества лет, потраченных на образование.

Этап 1

Так как образование родителей это две переменных (значит и два инструмента), то сначала проверим один инструмент, а именно образование матери. На первом этапе этого метода мы проверяем есть ли корреляция между mcollege и education. В Таблице 3 представлены результаты линейной регрессии, и если посмотреть на значение p-value (которое значимо), то мы сможем наблюдать достаточно сильную корреляцию между ними.

		Education	
Predictors	*Estimates*	*Std. Error*	*p*
(Intercept)	13.55	0.05	**<0.001**
mcollege [yes]	0.94	0.07	**<0.001**
gender [female]	0.02	0.05	0.657
ethnicity [afam]	-0.36	0.07	**<0.001**
ethnicity [hispanic]	-0.09	0.07	0.188
income [high]	0.65	0.06	**<0.001**
Observations	4739		
R^2 / R^2 adjusted	0.085 / 0.084		

Таблица 3. Первый этап. Результаты линейной регрессии для определения корреляции инструмента (mcollege) и эндогенной переменной (education).

Этап 2

Вторым этапом мы будем рассчитывать предсказанные значения для education (educationHat) и включать их в первую модель (см. предыдущую часть) вместо изначальных значений переменной education. В Таблице 4 мы видим новые коэффициенты, которые являются более точной оценкой эффекта переменной education.

		Score	
Predictors	*Estimates*	*Std. Error*	*p*
(Intercept)	1.61	5.03	0.749
educationHat	3.77	0.37	**<0.001**
gender [female]	-1.20	0.23	**<0.001**
ethnicity [afam]	-6.02	0.35	**<0.001**
ethnicity [hispanic]	-4.22	0.31	**<0.001**
income [high]	-0.86	0.40	**0.031**
Observations	4739		
R^2 / R^2 adjusted	0.085 / 0.084		

Таблица 4. Второй этап**.** Результаты линейной регрессии – регрессия такая же, как и для Таблицы 2, только вместо переменной education использовалась educationHat.

Другие инструменты

Однако образование родителей может по-разному коррелировать в зависимости от того, у кого есть высшее образование – у отца или матери, или у обоих. Чтобы оценить все эти три варианта, мы использовали тот же метод, что и выше, однако изменив инструмент (или добавив еще один). В Таблице 5 представлены результаты второго этапа трех МНК-регрессий: (1) инструмент – mcollege, (2) fcollege , (3) mcollege и fcollege. Как мы можем отметить, результаты в целом не отличаются друг от друга.

		Score	
Predictors	*mcollege*	*fcollege*	*mcollege u fcollege*
(Intercept)	1.61	2.81	2.40
	(5.03)	(4.08)	(3.67)
educationHat	**3.77*****	**3.68*****	**3.71*****
	(0.37)	(0.30)	(0.27)
gender [female]	**-1.20*****	**-1.20*****	**-1.20*****
	(0.23)	(0.23)	(0.23)
ethnicity [afam]	**-6.02*****	**-6.05*****	**-6.04*****
	(0.35)	(0.34)	(0.33)
ethnicity [hispanic]	**-4.22*****	**-4.23*****	**-4.23*****
	(0.31)	(0.31)	(0.30)
income [high]	**-0.86***	**-0.79***	**-0.81***
	(0.40)	(0.36)	(0.34)
Note:			*p<0.05; ** p<0.01; ***p<0.001

Таблица 5. В таблице представлены коэффициенты линейной регрессии. В скобках указаны стандартные ошибки.

Однако, как и было указано выше, стандартные ошибки с помощью метода МНК-регрессии рассчитываются не точно. Для того, чтобы уточнить результаты, используем другой способ расчета и попробуем сравнить результаты двух методов.

Ivreg

С помощью функции ivreg() мы получили точно такие же результаты, а изменения в стандартных ошибках почти не различимы (см. Таблицу 6).

		Score	
Predictors	*mcollege*	*fcollege*	*mcollege и fcollege*
(Intercept)	1.61	2.81	2.40
	(4.92)	(3.98)	(3.61)
education	**3.77*** **	**3.68*** **	**3.71*** **
	(0.36)	(0.29)	(0.26)
gender [female]	**-1.20*** **	**-1.20*** **	**-1.20*** **
	(0.23)	(0.23)	(0.23)
ethnicity [afam]	**-6.02*** **	**-6.05*** **	**-6.04*** **
	(0.34)	(0.33)	(0.33)
ethnicity [hispanic]	**-4.22*** **	**-4.23*** **	**-4.23*** **
	(0.30)	(0.30)	(0.30)
income [high]	**-0.86***	**-0.79***	**-0.81***
	(0.39)	(0.35)	(0.33)
Note:			*p<0.05; ** p<0.01; ***p<0.001

Таблица 6. Результаты функций ivreg. В таблице представлены коэффициенты. В скобках указаны стандартные ошибки.

Сравнение

Для сравнения результатов двух методов мы будет использовать модели с двумя инструментами – образования матери и отца. А также сравним данные результаты с первой моделью, которую мы рассчитывали с использованием линейной регрессии. В Таблице 7 представлены результаты.

		Score	
Predictors	*Без инструментов*	*МНК*	*Ivreg*
(Intercept)	24.85	2.40	2.40
	(0.84)	(3.67)	(3.61)
education	**2.06*** **	**3.71*** **	**3.71*** **
	(0.06)	(0.27)	(0.26)
gender [female]	**-1.17*** **	**-1.20*** **	**-1.20*** **
	(0.21)	(0.23)	(0.23)
ethnicity [afam]	**-6.65*** **	**-6.04*** **	**-6.04*** **
	(0.29)	(0.33)	(0.33)
ethnicity [hispanic]	**-4.48*** **	**-4.23*** **	**-4.23*** **
	(0.28)	(0.30)	(0.30)
income [high]	**0.54***	**-0.81***	**-0.81***
	(0.24)	(0.34)	(0.33)
Note:			*p<0.05; ** p<0.01; ***p<0.001

Таблица 7. Результаты первоначальной модели и после метода инструментальных переменных (МНК и Ivreg). В таблице представлены коэффициенты. В скобках указаны стандартные ошибки.

Если посмотреть на эндогенную переменную education (количество лет на образование), то ее значимость увеличивается при использовании инструментальных переменных mcollege и fcollege. Причем, как и было отмечено ранее, результаты расчета эффекта как при использовании МНК-регрессии, так и ivreg() не отличаются. Также стоит отметить, что изменил свой полюс эффект дохода семьи – изначально он был положительный, а стал отрицательным.

Оценка силы инструмента

Для проверки силы инструмента использовался F-test, а в качестве модели - результаты ivreg метода. Так как результаты теста значимы (p-value = 0.00), то мы можем сказать, что наши инструментальные переменные являются сильными (см. Таблицу 8).

Тест Ву-Хаусмана на экзогенность показывает значимые результаты (p-value = 0.00), следовательно, переменная education является эндогенной. Отсюда мы можем заключить, что необходимость использования инструмента оправдана и она показывает более точные результаты, чем обычная МНК-регрессия.

С помощью теста Саргана мы проверяем, что все инструменты (а у нас их два) являются валидными. Результаты теста оказываются незначимыми (p-value = 0.811), поэтому мы можем заключить, что оба инструмента являются валидным.

	mcollege	*fcollege*	*mcollege и fcollege*
F-test	**158.725**	**242.705**	**150.477**
	(0)	**(0)**	**(0)**
Wu-Hausman	27.255	37.559	48.441
	(0)	(0)	(0)
Sargan	NA	NA	0.057
			(0.811)

Таблица 8. Результаты F-test, Wu-Hausman test, Sargan test. Значения p-value указаны в скобках.

Вывод

Главным выводом данной работы является заключение о том, что использование метода инструментальных переменных будет куда более точным для определения того, что и как влияет на успеваемость студента. Как было показано выше, эффект от количества лет, затраченных на образование, является куда более сильным, чем было рассчитано изначально. Т.е. если бы мы оценивали этот эффект без использования метода инструментальных переменных, то могли бы недооценить его. Более того, было обнаружено, что и доход семьи (при использовании инструментов) оказывает не положительный, а даже негативный эффект на успеваемость.

Также стоит отметить, что использование в качестве инструментов образования родителей не варьировалась от отца к матери, что наводит нас на заключение о том, что вне зависимости от того, у кого в семье есть высшее образование, эффект будет одинаков для студента.