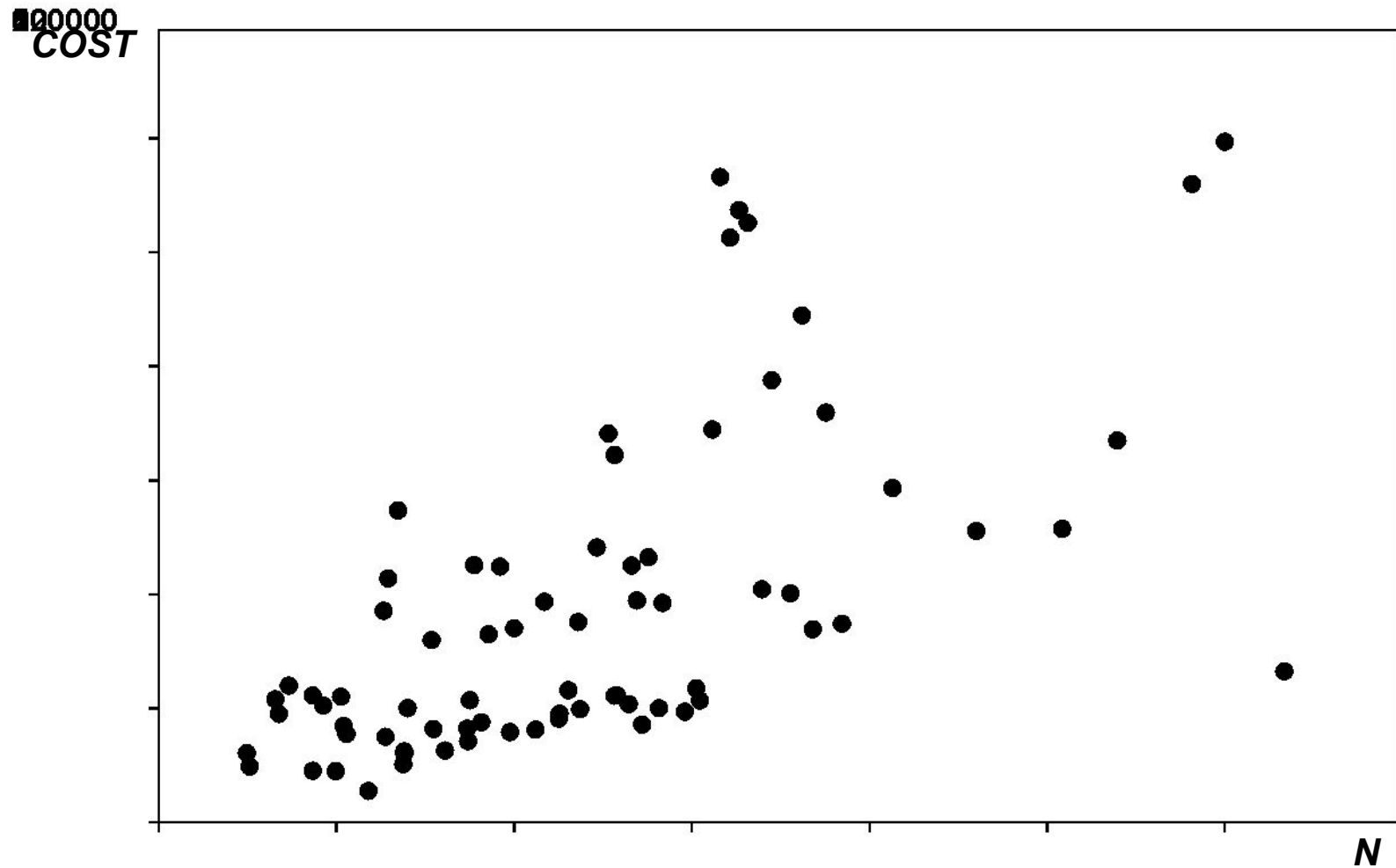
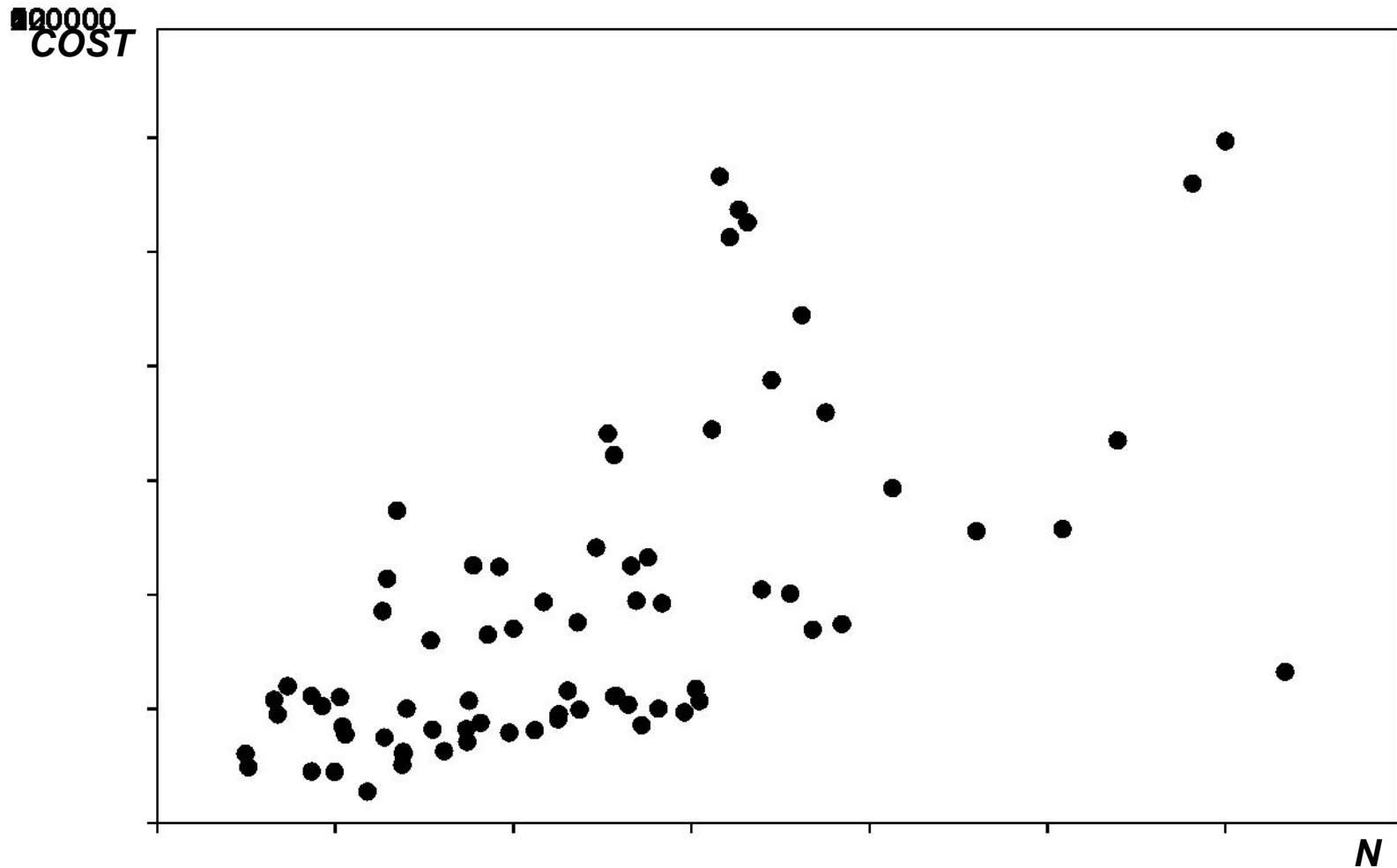


ТЕСТ ЧОУ

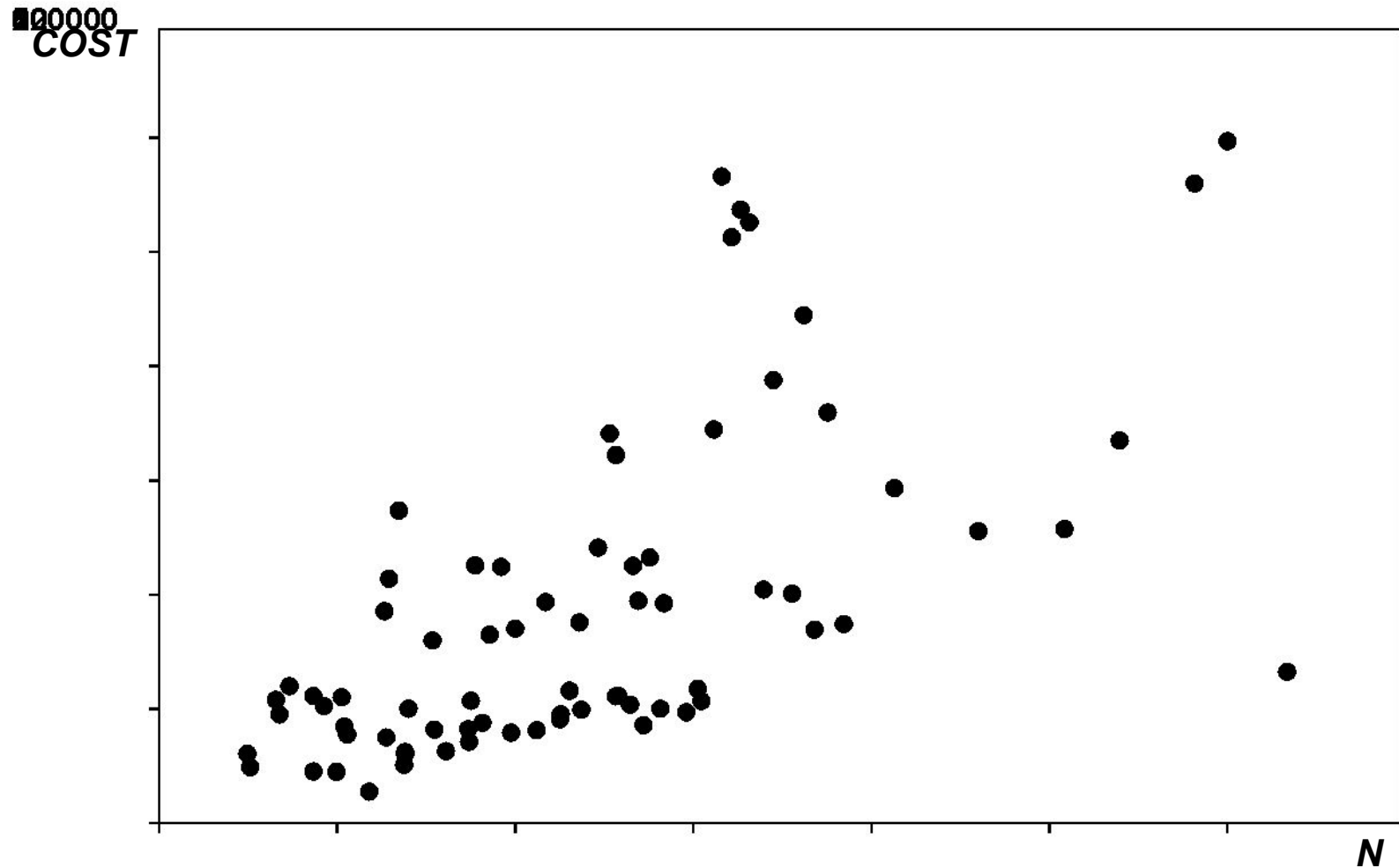


Иногда при регрессионном анализе в образце данных имеются два типа наблюдений.

ТЕСТ ЧОУ



Если это так, разумно исследовать, применима ли одна модель регрессии к обоим категориям или нужны ли вам отдельные для них. Для этого вы можете выполнить тест Chow



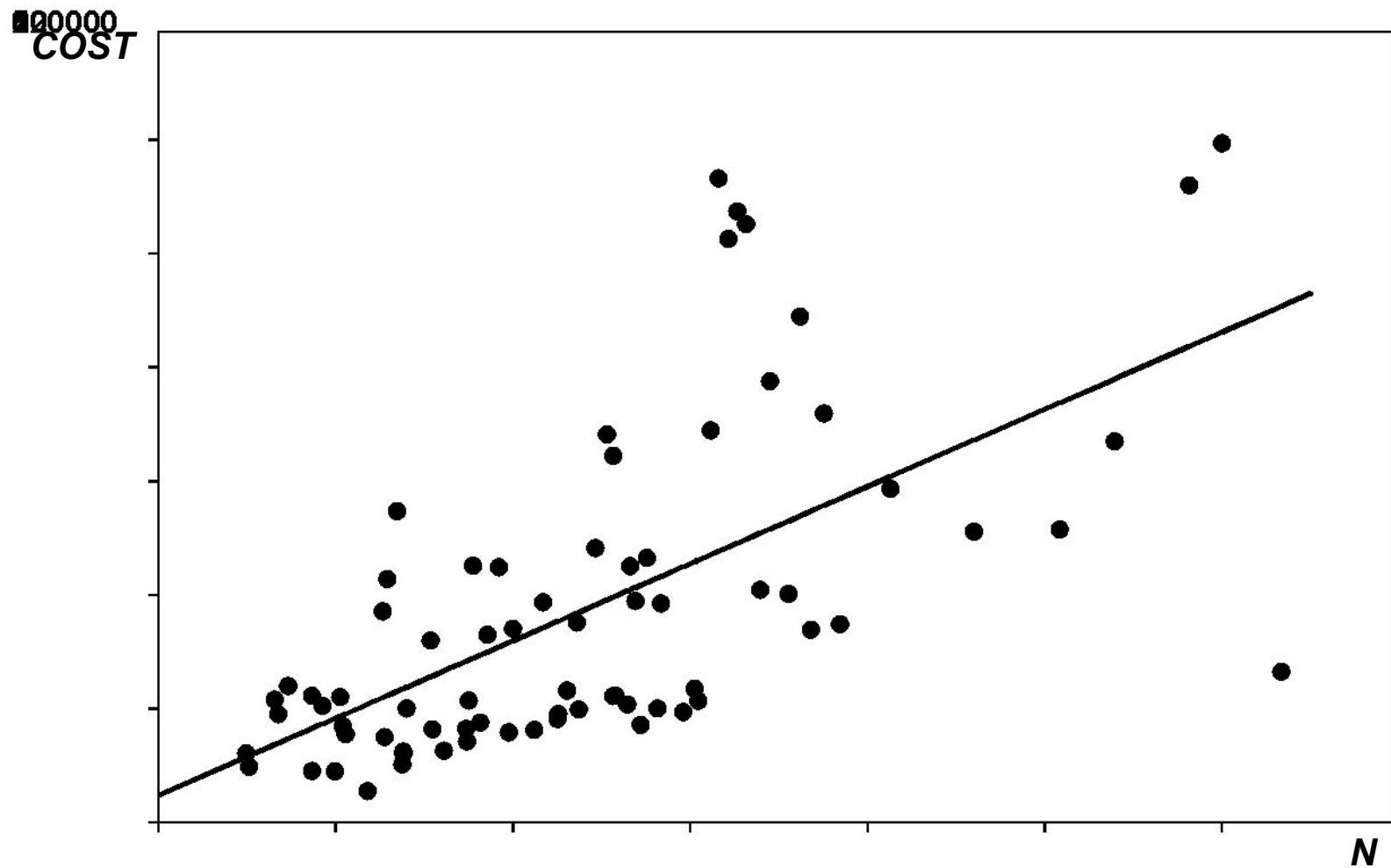
Мы проиллюстрируем это, используя данные для 74 средних школ в Шанхае. Диаграмма разброса отображает данные о годовых периодических расходах и числе студентов.

```
. reg COST N
```

Source	SS	df	MS	Number of obs = 74		
Model	5.7974e+11	1	5.7974e+11	F(1, 72)	=	46.82
Residual	8.9160e+11	72	1.2383e+10	Prob > F	=	0.0000
Total	1.4713e+12	73	2.0155e+10	R-squared	=	0.3940
				Adj R-squared	=	0.3856
				Root MSE	=	1.1e+05

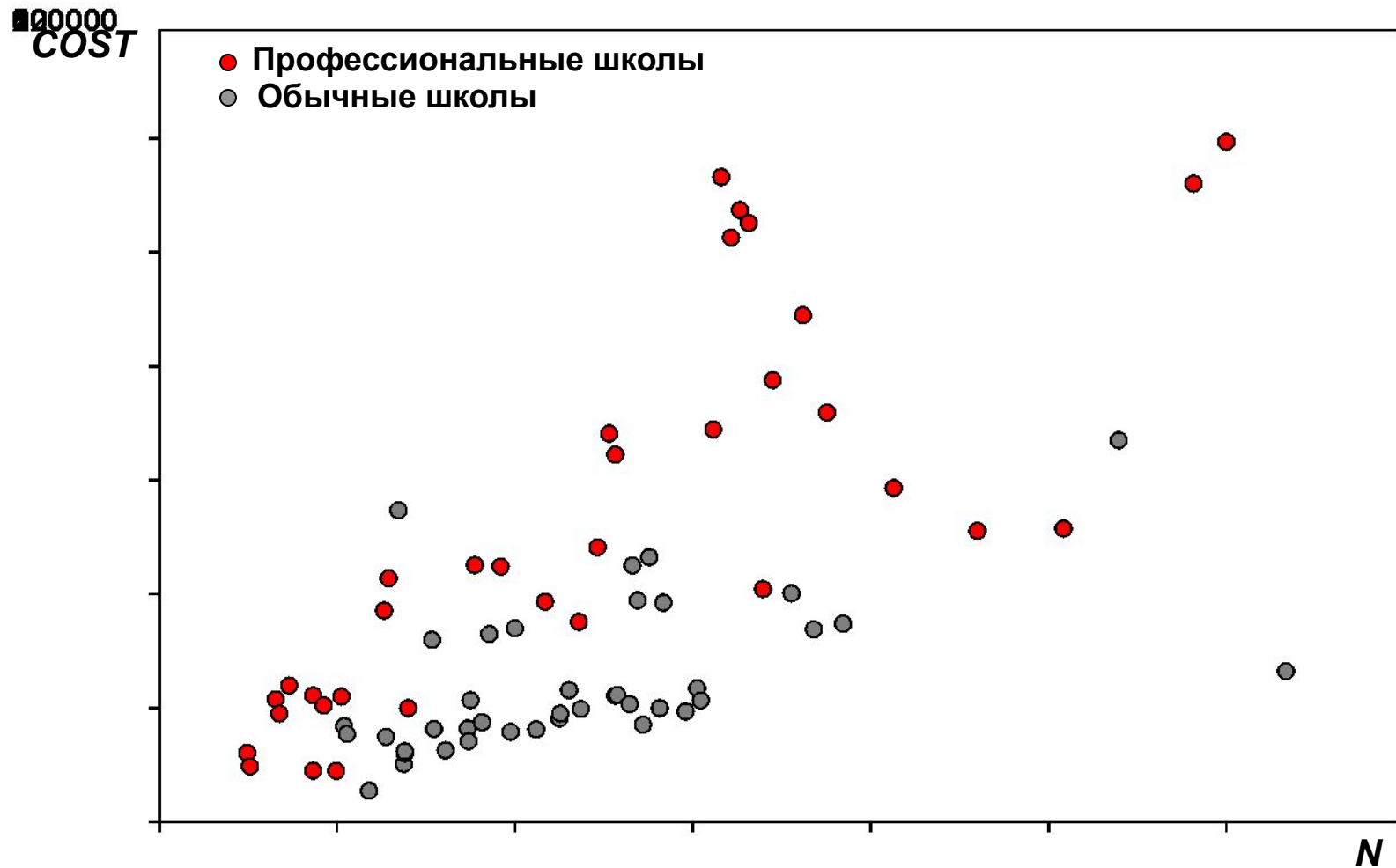
COST	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
N	339.0432	49.55144	6.842	0.000	240.2642	437.8222
_cons	23953.3	27167.96	0.882	0.381	-30205.04	78111.65

Вот результат регрессии, когда COST регрессируется на N, не делая различия между различными типами школ.



Это диаграмма рассеяния с линией регрессии

ТЕСТ ЧОУ



Теперь мы проводим различие между профессиональными школами и обычными школами и проводим отдельные регрессии для двух подвыборки

```
. reg COST N if OCC==1
```

Source	SS	df	MS	Number of obs = 34		
Model	6.0538e+11	1	6.0538e+11	F(1, 32)	=	55.52
Residual	3.4895e+11	32	1.0905e+10	Prob > F	=	0.0000
Total	9.5433e+11	33	2.8919e+10	R-squared	=	0.6344
				Adj R-squared	=	0.6229
				Root MSE	=	1.0e+05

COST	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
N	436.7769	58.62085	7.451	0.000	317.3701	556.1836
_cons	47974.07	33879.03	1.416	0.166	-21035.26	116983.4

Это результат регрессии, когда COST регрессируется на N, используя подвыборку из 34 профессиональных школ.

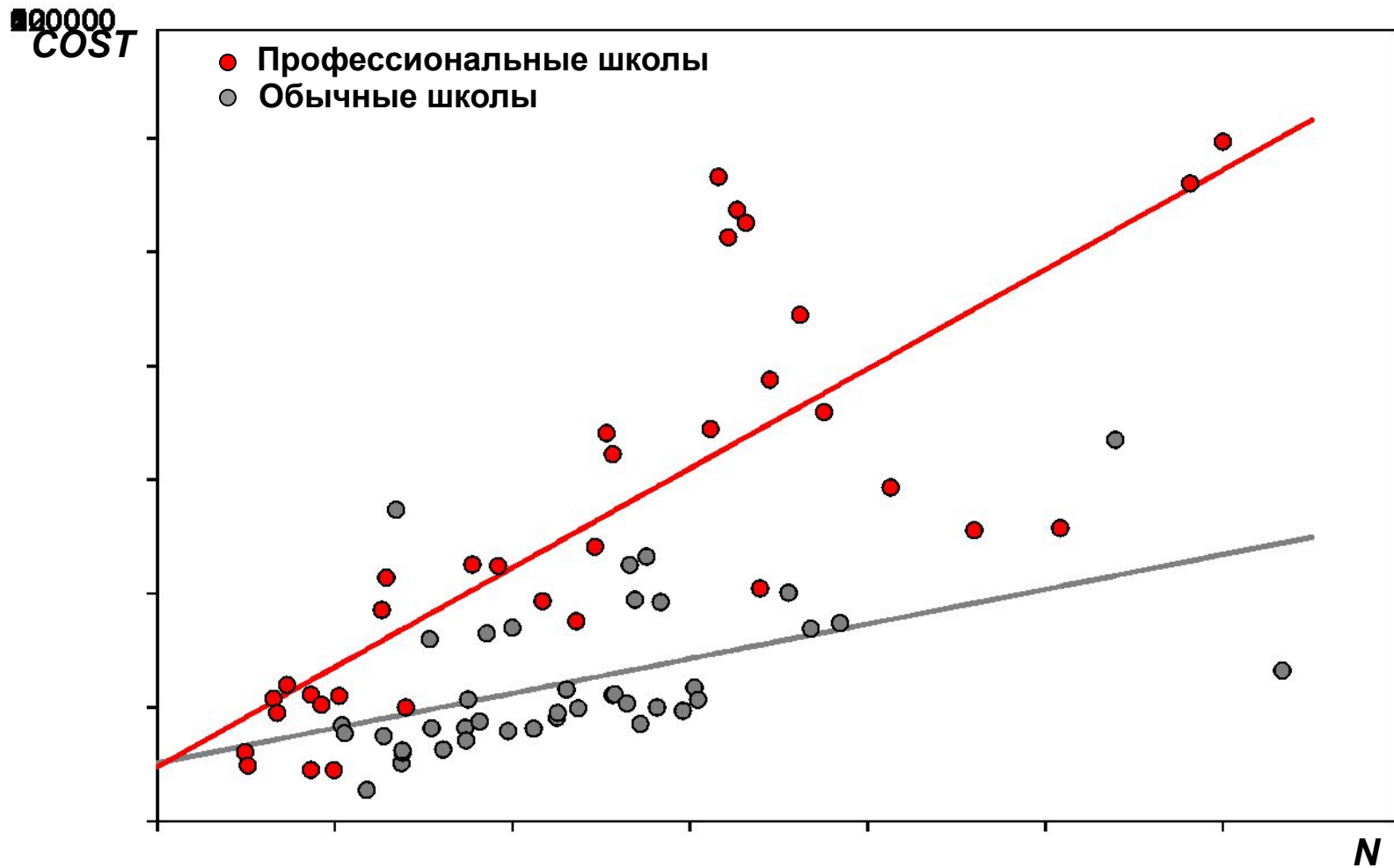
```
. reg COST N if OCC==0
```

Source	SS	df	MS	Number of obs = 40		
Model	4.3273e+10	1	4.3273e+10	F(1, 38)	=	13.53
Residual	1.2150e+11	38	3.1973e+09	Prob > F	=	0.0007
Total	1.6477e+11	39	4.2249e+09	R-squared	=	0.2626
				Adj R-squared	=	0.2432
				Root MSE	=	56545

COST	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
N	152.2982	41.39782	3.679	0.001	68.49275	236.1037
_cons	51475.25	21599.14	2.383	0.022	7750.064	95200.43

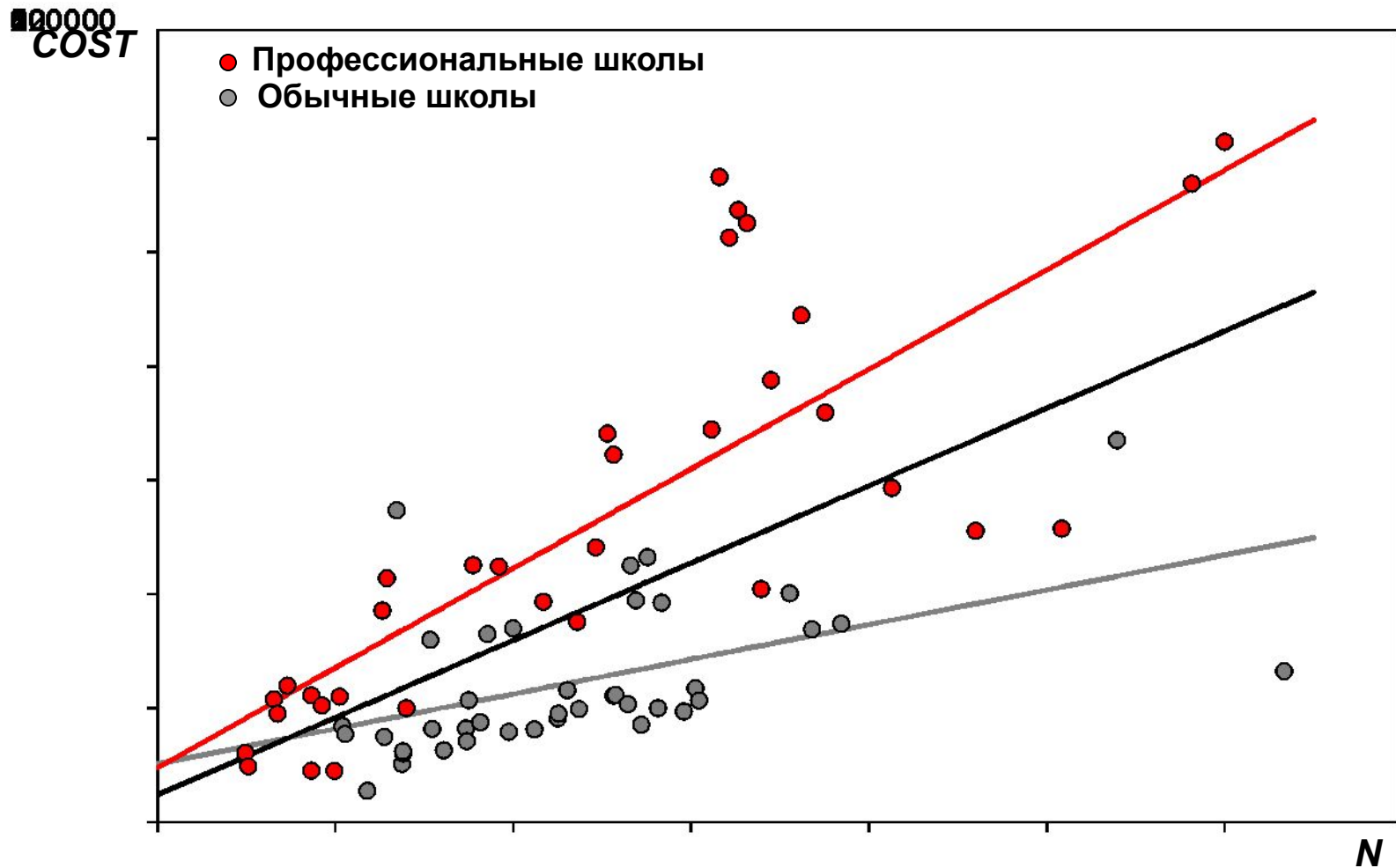
И это результат регрессии, когда COST регрессируется на N для подвыборки из 40 обычных школ.

ТЕСТ ЧОУ



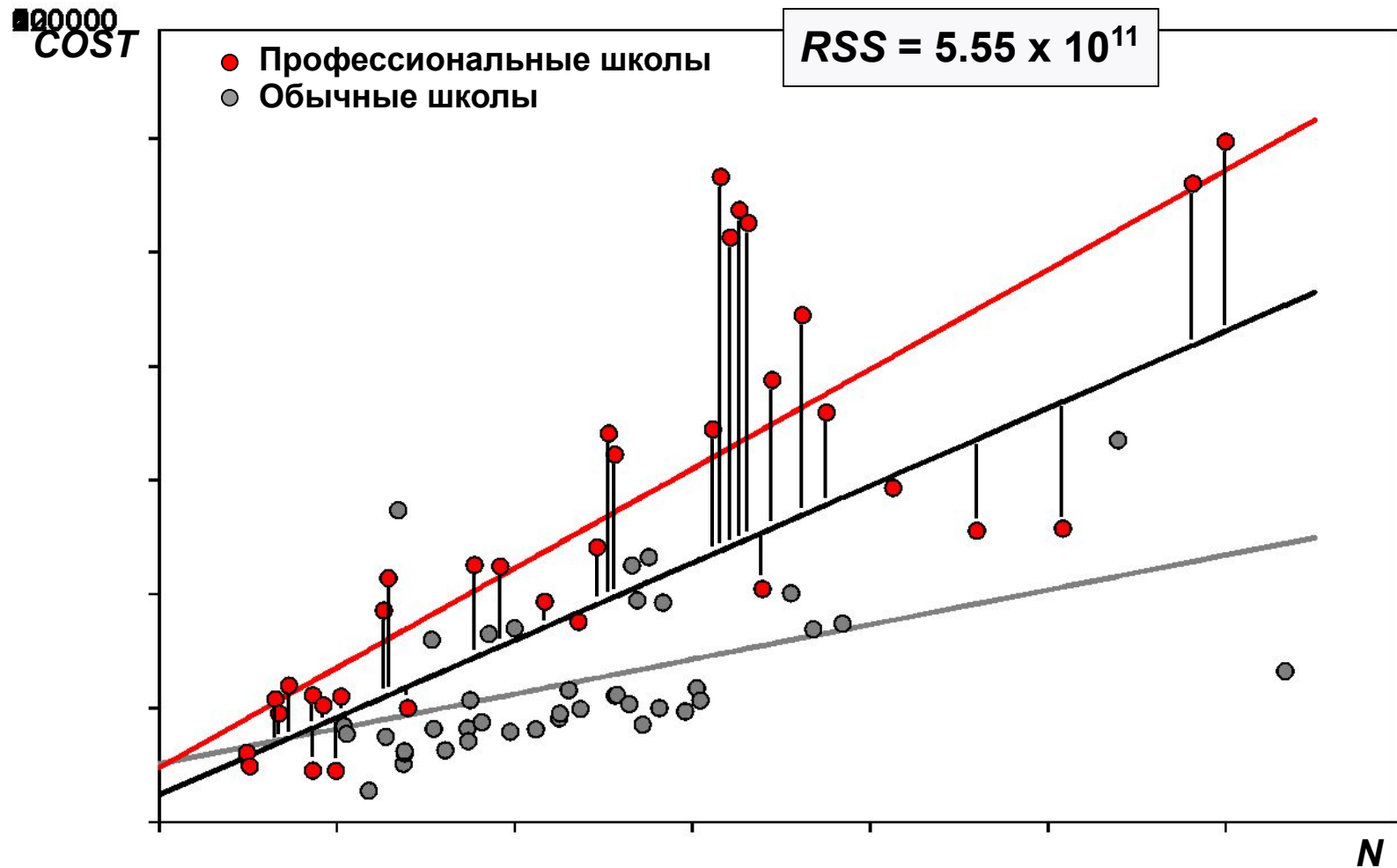
Вот регрессионные линии для двух подвыборки

ТЕСТ ЧОУ



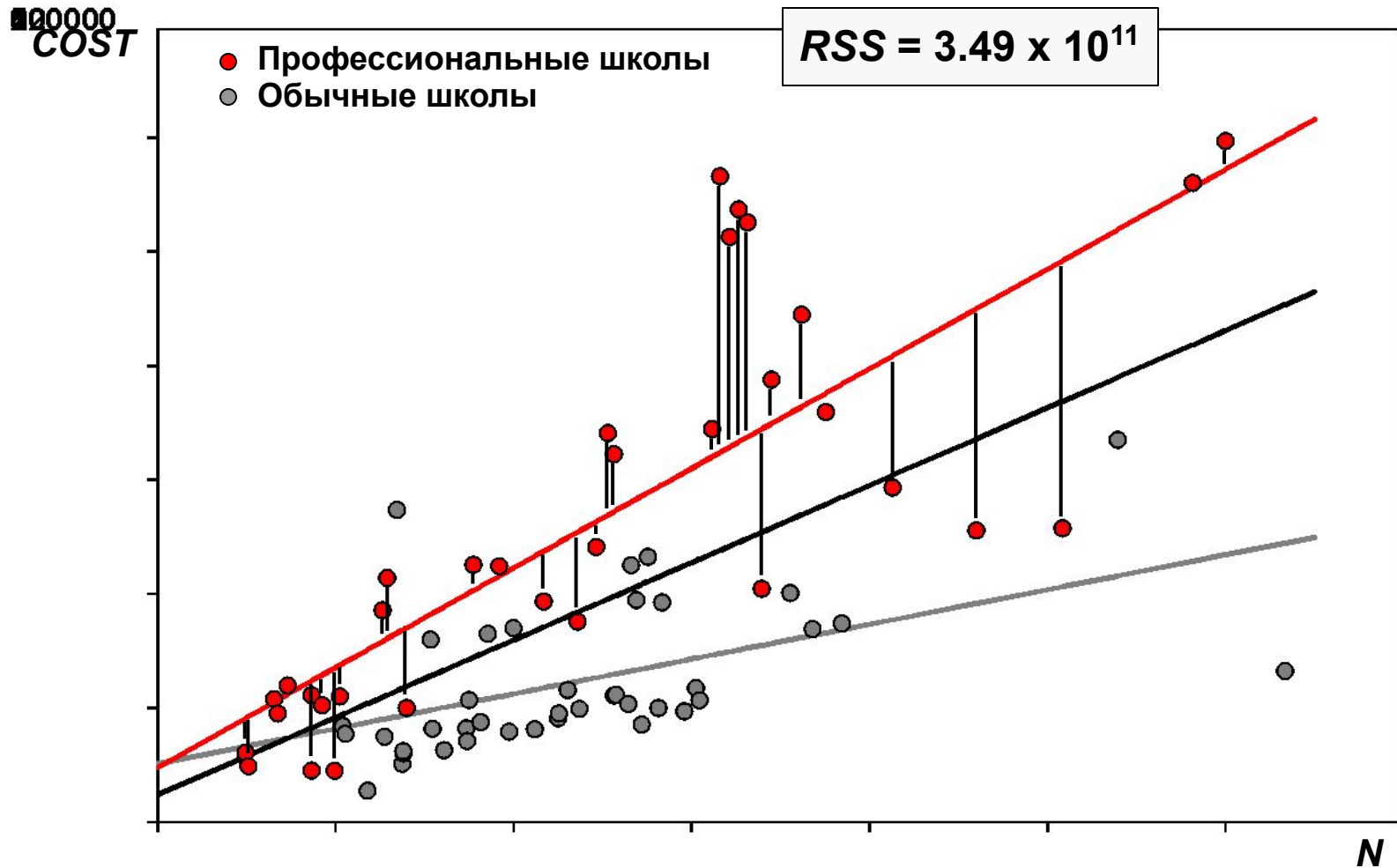
Для сравнения показана линия регрессии для объединенного образца (целая выборка, без какого-либо различия).

ТЕСТ ЧОУ



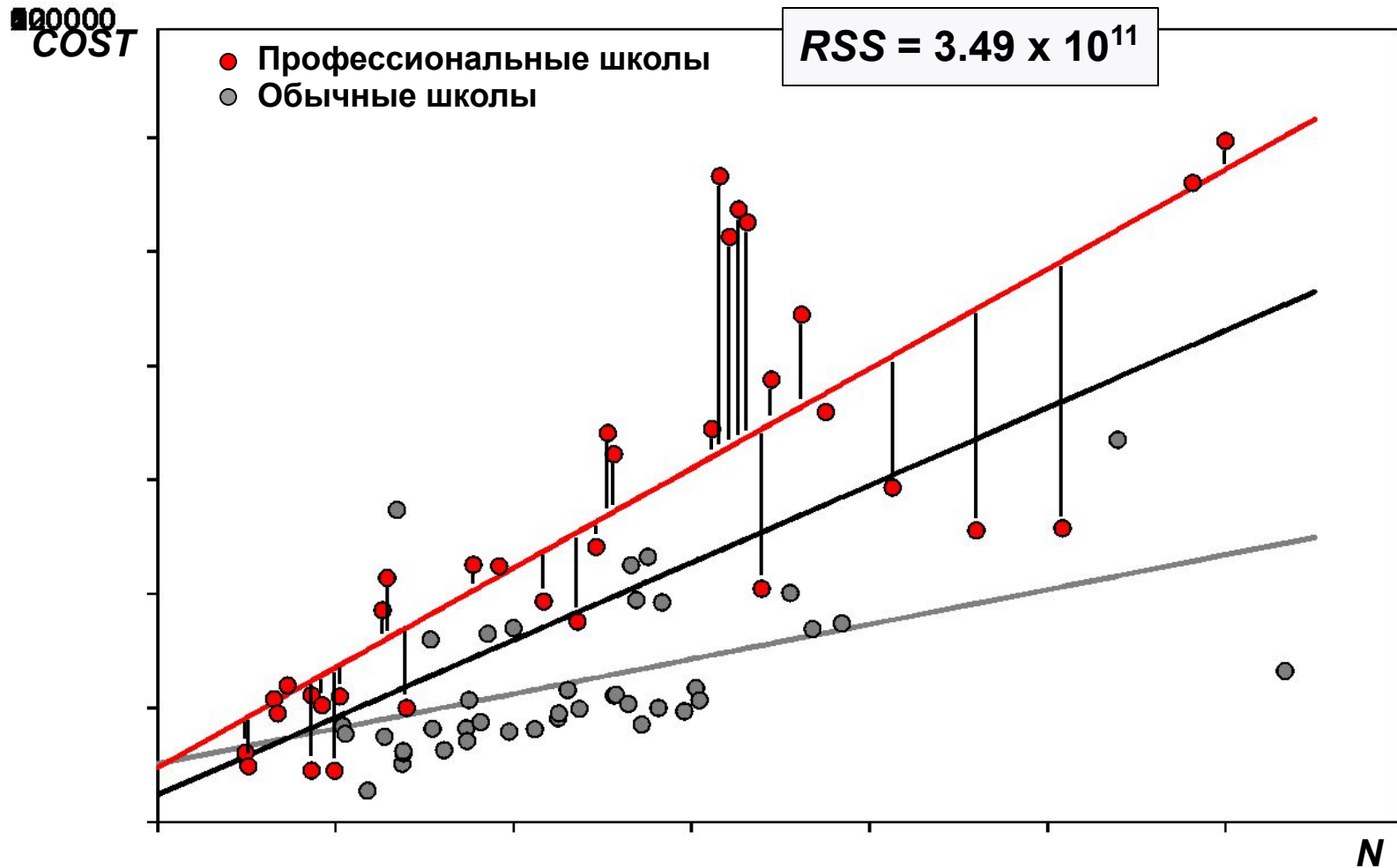
На диаграмме показаны остатки для профессиональных школ в регрессии с использованием объединенного образца

ТЕСТ ЧОУ

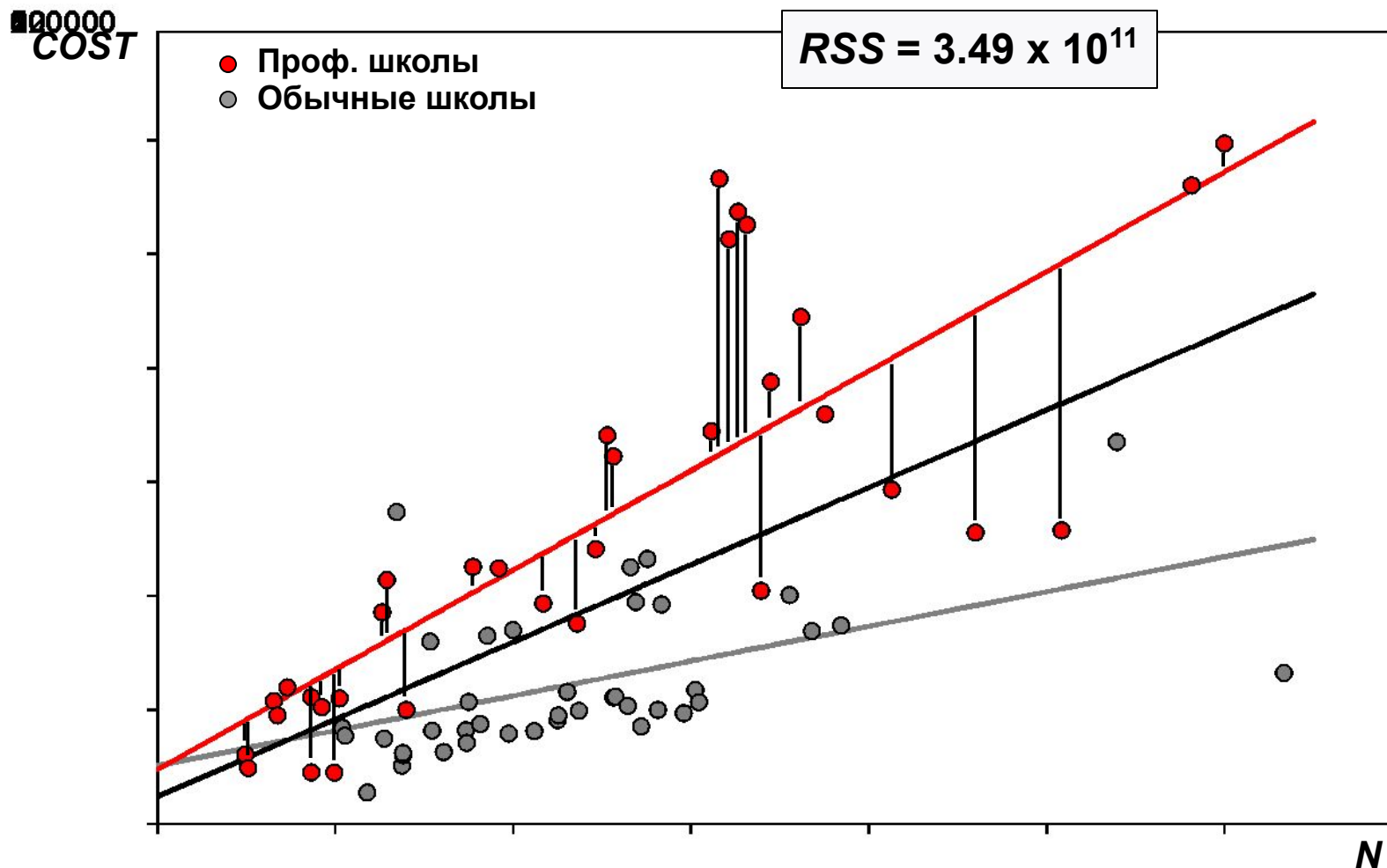


Теперь соответствующие остатки для регрессии с использованием только выборочных наблюдений в оккупационных школах

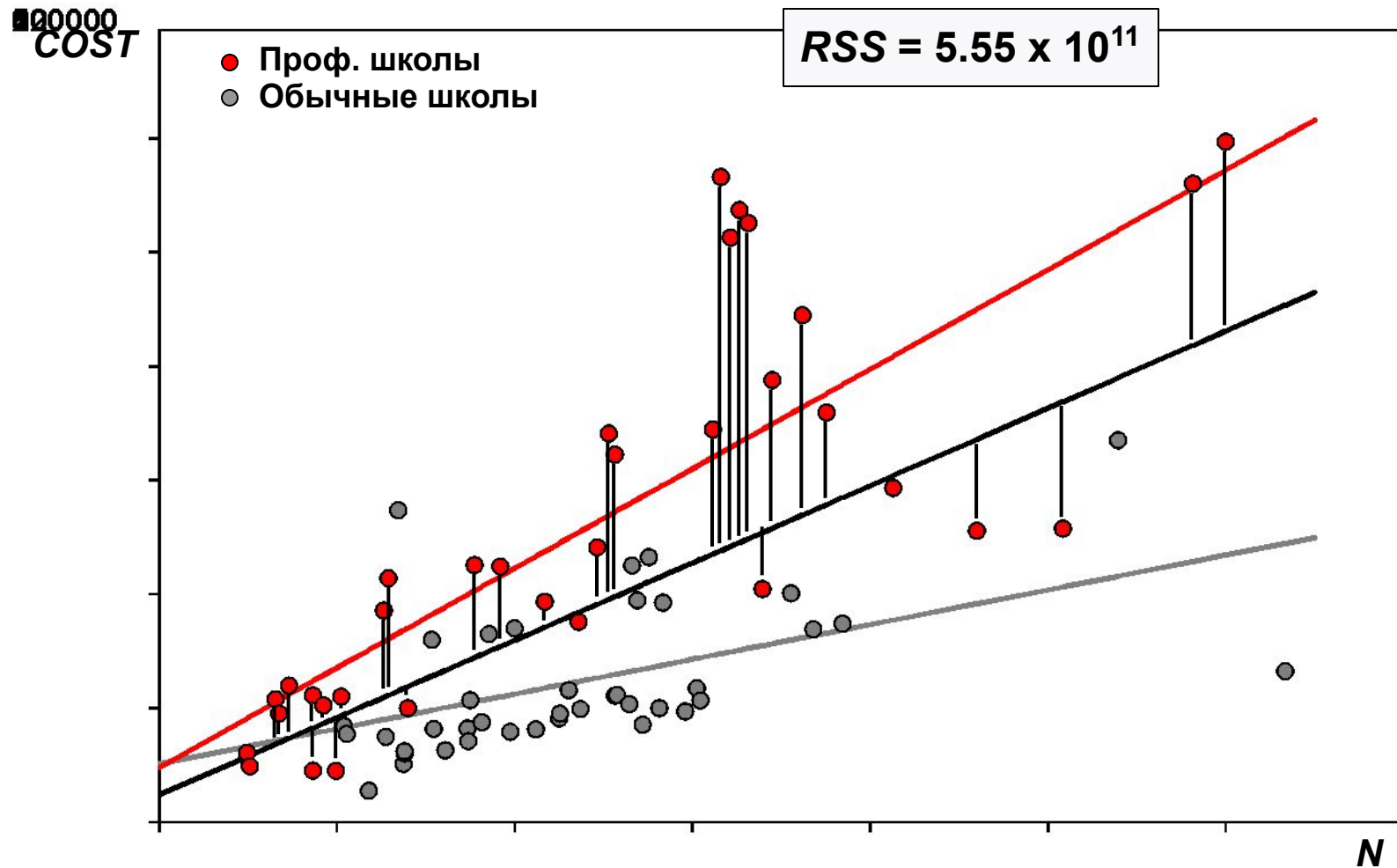
ТЕСТ ЧОУ



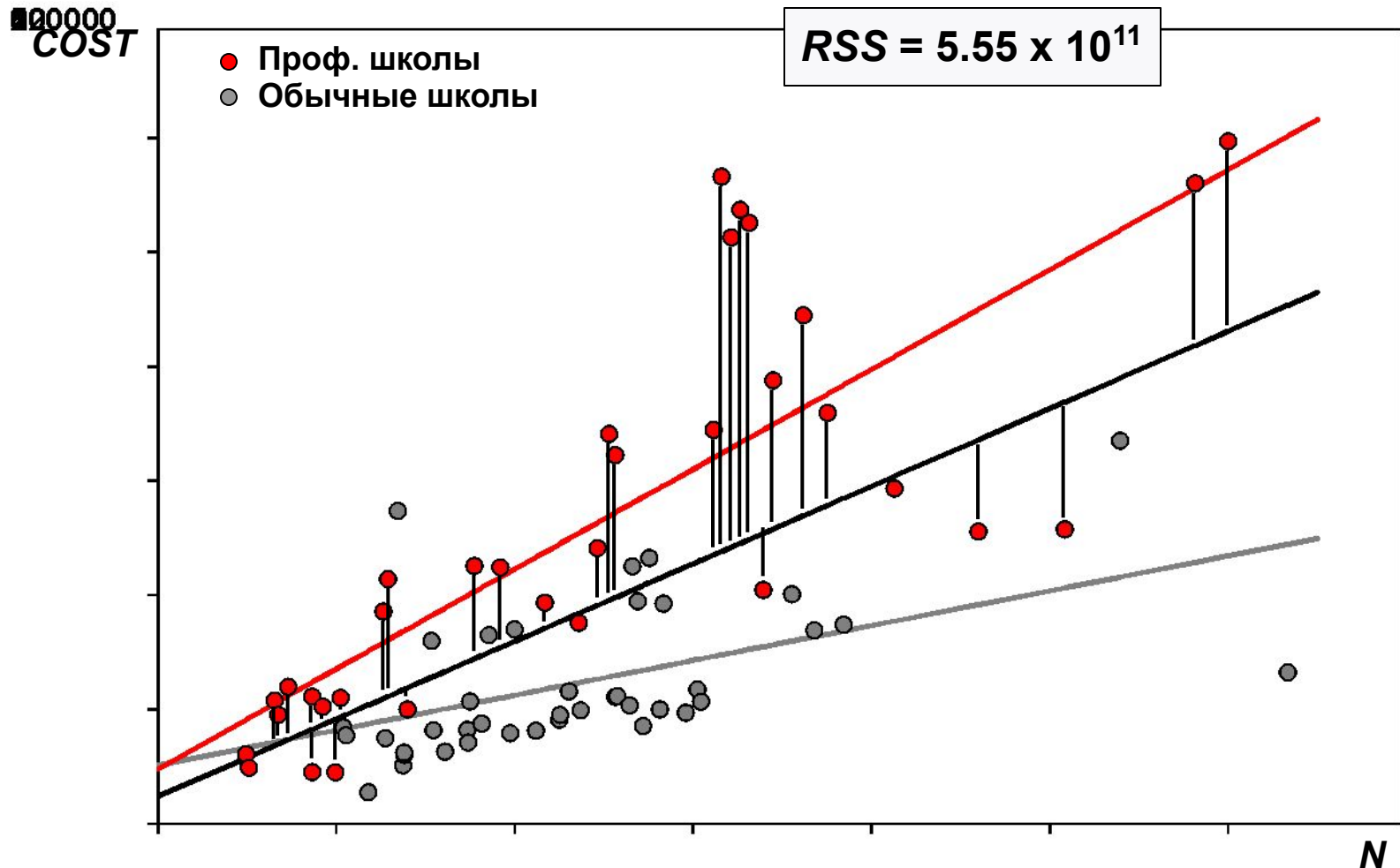
RSS меньше для остатков от подвыборной регрессии. Это должно быть так. Зачем? (Попробуйте ответить, прежде чем продолжить.)



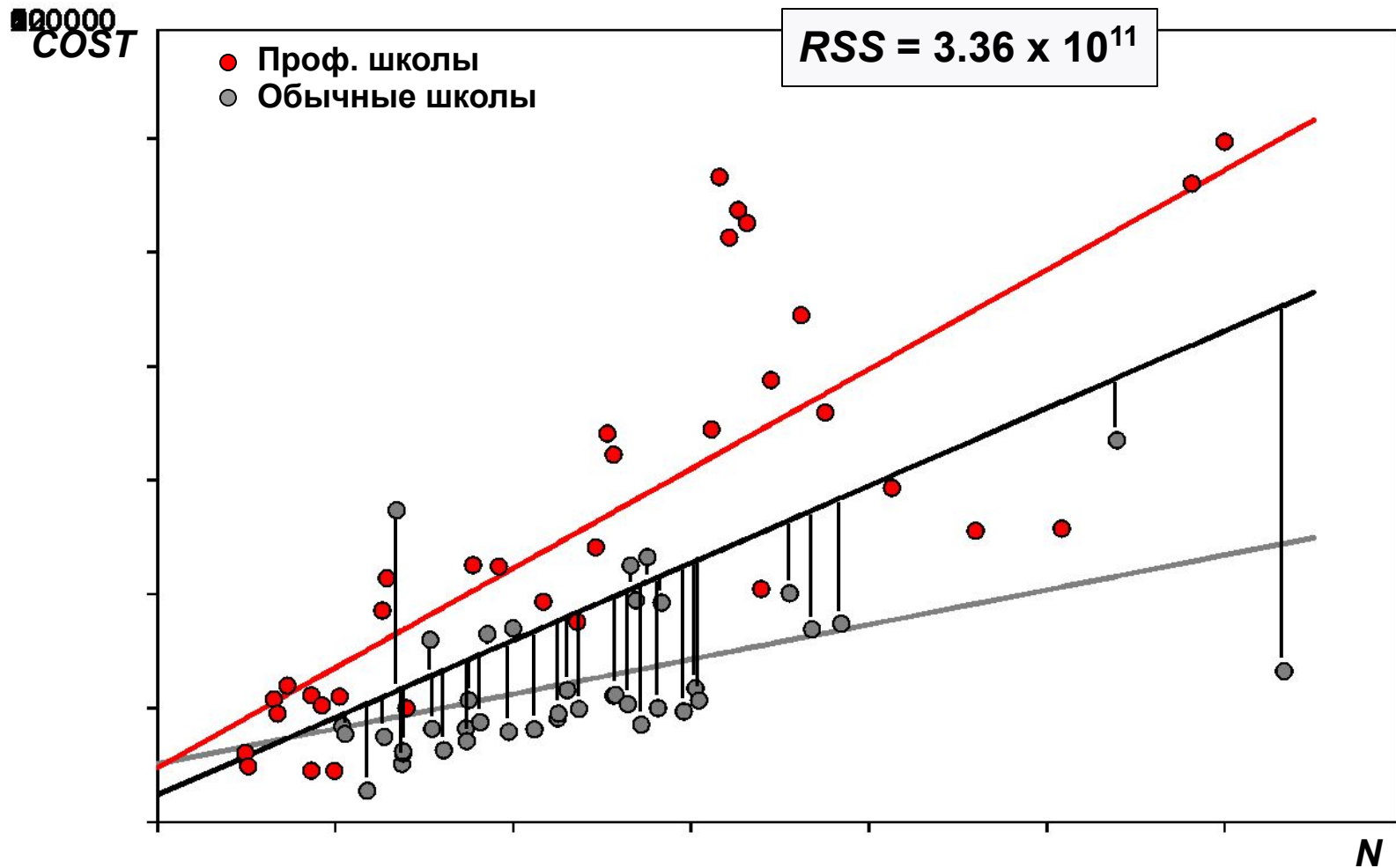
Линия регрессии для подвыборной регрессии расположена так, чтобы минимизировать сумму квадратов остатков для наблюдений в школе. Это принцип, лежащий в основе OLS.



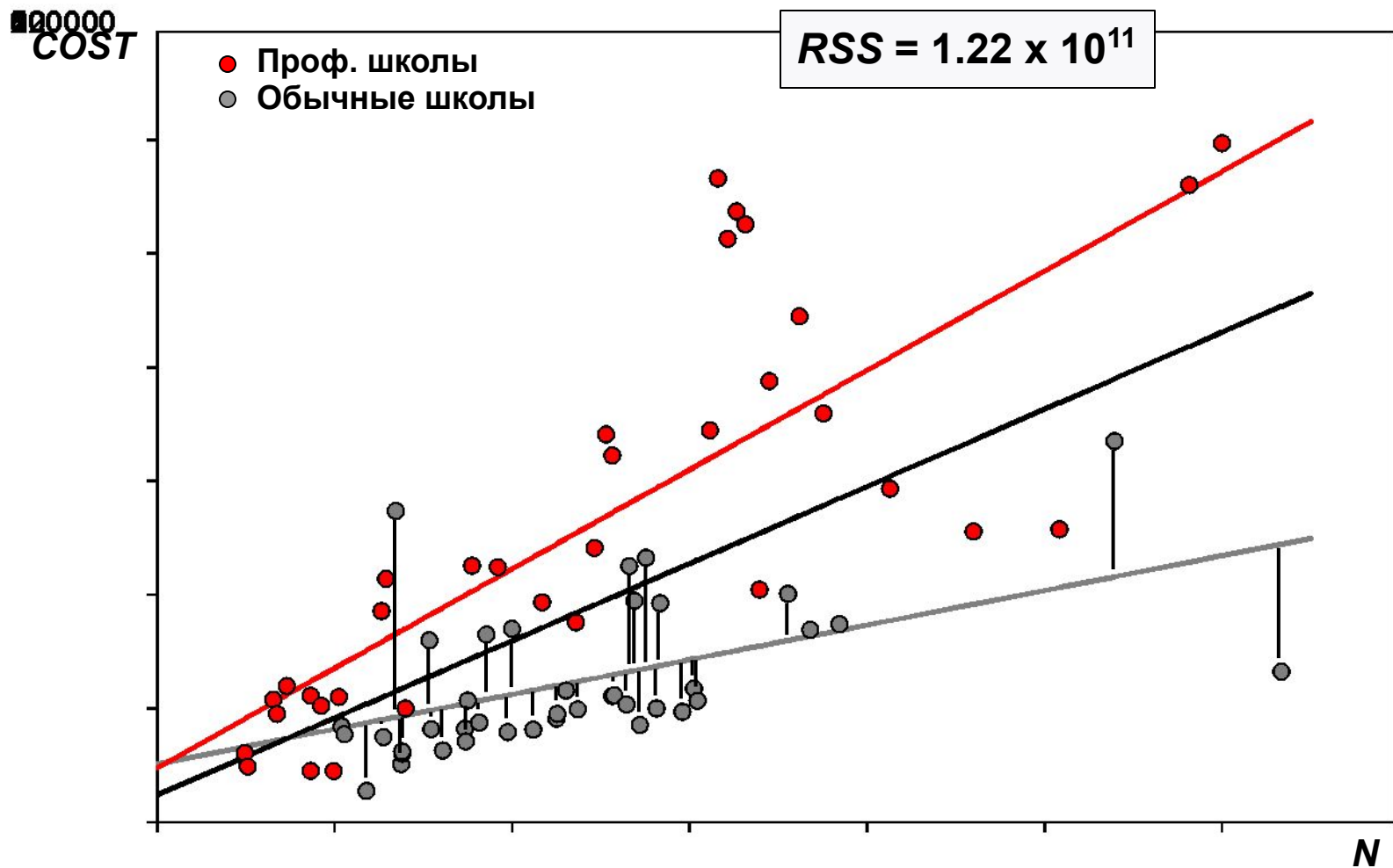
Линия регрессии для объединенного образца расположена для обеспечения наилучшего общего соответствия для образца в целом, включая обычные школы



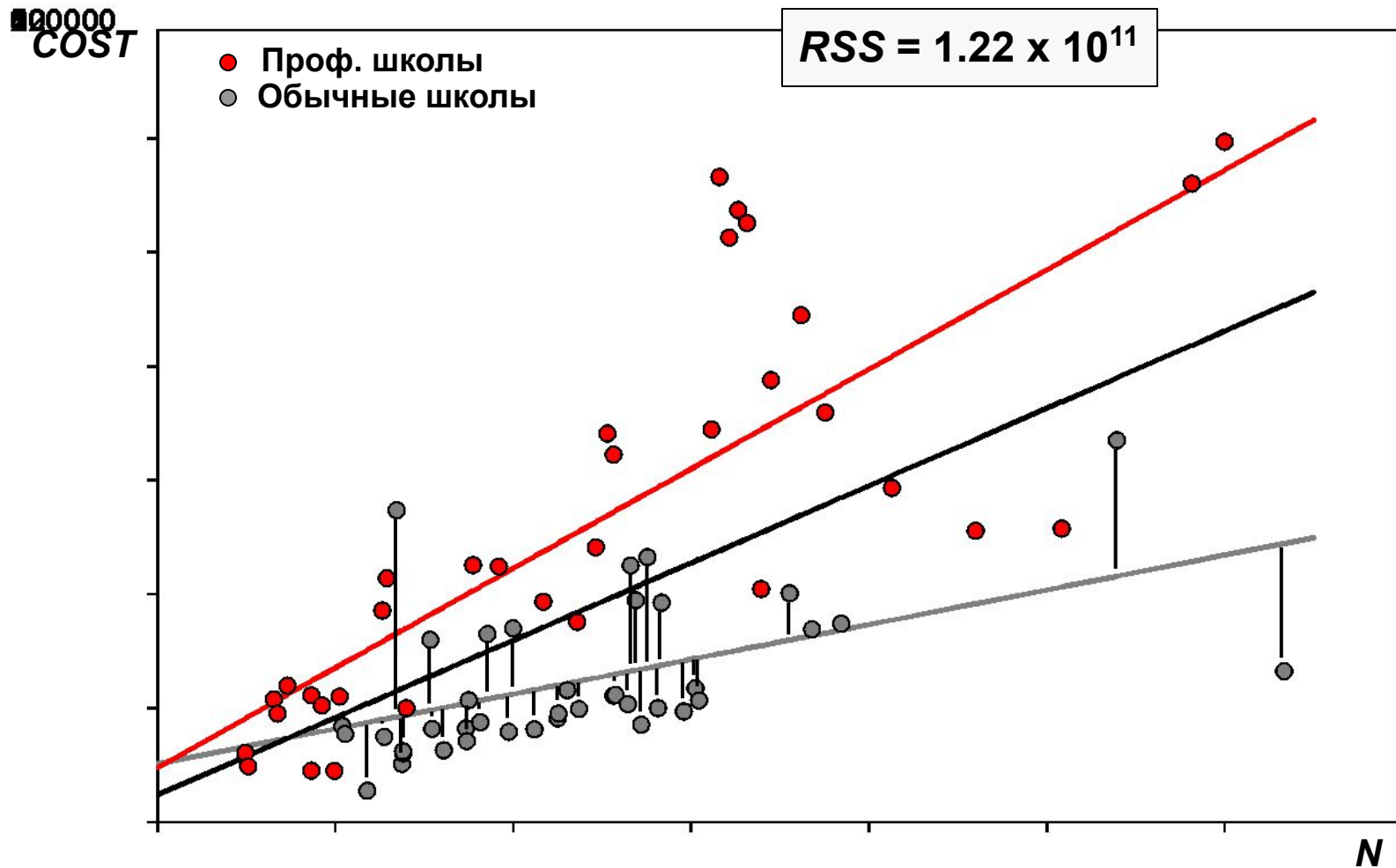
Поэтому его местоположение является компромиссом между наилучшим подходом к наблюдениям в профессиональной школе и наилучшим образом подходит для обычных школьных наблюдений. Поскольку это компромисс, его подгонка будет ниже, чем для регрессии подвыборки



Затем мы переходим к обычным школам. Вот остатки для объединенной регрессии.



Затем мы переходим к обычным школам. Вот остатки для объединенной регрессии



Опять же, RSS должен быть ниже для регрессии подвыборки, чем для объединенной регрессии выборки.

Остаточная сумма квадратов ($\times 10^{11}$)

Регрессия	Профессиональная	Обычная	Всего
Отдельный	$RSS_1 = 3.49$	$RSS_2 = 1.22$	4.71
Объединенный	5.55	3.36	$RSS_p = 8.91$

В таблице приведены данные RSS для двух типов школ в отдельных и объединенных регрессиях.

Остаточная сумма квадратов($\times 10^{11}$)

Регрессия	Профессиональная	Обычная	Всего
Отдельный $RSS_1 = 3.49$	$RSS_2 = 1.22$	4.71	
Объединенный	5.55	3.36	$RSS_p = 8.91$

Остаточные суммы квадратов для отдельных регрессий для профессиональных и обычных школ будут обозначаться соответственно RSS_1 и RSS_2

Остаточная сумма квадратов ($\times 10^{11}$)

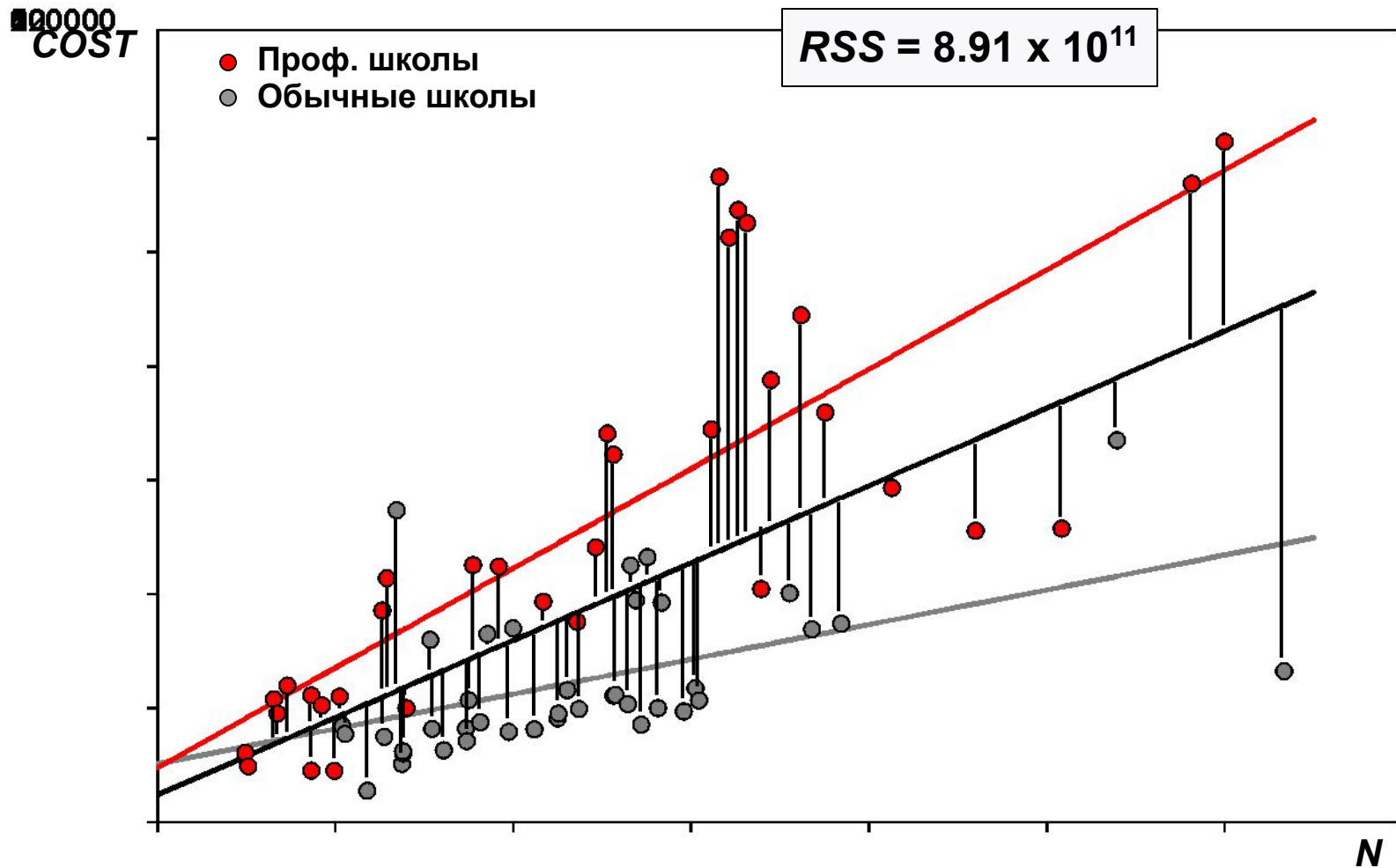
Регрессия	Профессиональная	Обычная	Всего
Отдельные $RSS_1 = 3.49$	$RSS_2 = 1.22$	4.71	
Объединенные	5.55	3.36	$RSS_p = 8.91$

Добавляя их вместе, мы получаем общую остаточную сумму квадратов, когда для двух подвыборки выполняются отдельные регрессии.

Остаточная сумма квадратов($\times 10^{11}$)

Регрессия	Профессиональный	Обычная	Всего
Отдельные	$RSS_1 = 3.49$	$RSS_2 = 1.22$	4.71
Объединенные	5.55	3.36	$RSS_p = 8.91$

Мы сравниваем эту сумму с RSS_p , остаточной суммой квадратов из объединенной регрессии выборки.



Это получается непосредственно из исходной регрессии. Нет необходимости вычислять профессиональные и регулярные компоненты. Нас интересует только итог.

Остаточная сумма квадратов ($\times 10^{11}$)

<i>Регрессия</i>	Профессиональная	Обычная	Всего
Отдельные $RSS_1 = 3.49$	$RSS_2 = 1.22$	4.71	
Объединенные 5.55	3.36	$RSS_p = 8.91$	

Мы заинтересованы в том, чтобы добиться существенного уменьшения общего количества, когда мы выполняем отдельные регрессии для двух подвыборки.

$$F(k, n - 2k) = \frac{\begin{array}{l} \text{общее снижение в RSS, когда} \\ \text{выполняются отдельные регрессии} \end{array} / \begin{array}{l} \text{стоимость в} \\ \text{степени} \\ \text{свободы} \end{array}}{\begin{array}{l} \text{общее количество оставшихся RSS} \\ \text{выполняются отдельные регрессии} \end{array} / \begin{array}{l} \text{степеней} \\ \text{свободы} \\ \text{осталось} \end{array}}$$

$$= \frac{(RSS_p - [RSS_1 + RSS_2]) / k}{(RSS_1 + RSS_2) / (n - 2k)}$$

Статистика теста - это статистика F, определенная как показано

$$F(k, n - 2k) = \frac{\text{общее снижение в RSS, когда выполняются отдельные регрессии}}{\text{стоимость в степени свободы}} \bigg/ \frac{\text{общее количество оставшихся RSS выполняются отдельные регрессии}}{\text{степеней свободы осталось}}$$

$$= \frac{(RSS_p - [RSS_1 + RSS_2]) / k}{(RSS_1 + RSS_2) / (n - 2k)}$$

Первым аргументом статистики F является k, стоимость, с точки зрения степеней свободы, выполнения отдельных регрессий.

$$F(k, n - 2k) = \frac{\text{общее снижение в RSS, когда выполняются отдельные регрессии}}{\text{общее количество оставшихся RSS выполняются отдельные регрессии}} \cdot \frac{\text{стоимость в степени свободы}}{\text{степеней свободы осталось}}$$

$$= \frac{(RSS_p - [RSS_1 + RSS_2]) / k}{(RSS_1 + RSS_2) / (n - 2k)}$$

Стоимость k , поскольку два набора k параметров оцениваются при выполнении отдельных регрессий, а не только один набор с объединенной регрессией.

$$F(k, n - 2k) = \frac{\text{общее снижение в RSS, когда выполняются отдельные регрессии}}{\text{стоимость в степени свободы}} \bigg/ \frac{\text{общее количество оставшихся RSS выполняются отдельные регрессии}}{\text{степеней свободы осталось}}$$

$$= \frac{(RSS_p - [RSS_1 + RSS_2]) / k}{(RSS_1 + RSS_2) / (n - 2k)}$$

Второй аргумент статистики F - $n - 2k$, общее число степеней свободы, оставшихся при выполнении отдельных регрессий.

$$F(k, n - 2k) = \frac{\text{общее снижение в RSS, когда выполняются отдельные регрессии}}{\text{стоимость в степени свободы}} = \frac{\text{общее количество оставшихся RSS выполняются отдельные регрессии}}{\text{степеней свободы осталось}} = \frac{(RSS_p - [RSS_1 + RSS_2]) / k}{(RSS_1 + RSS_2) / (n - 2k)}$$

По каждой регрессии, когда проводятся отдельные регрессии, есть n наблюдений и k степеней свободы.

$$F(k, n - 2k) = \frac{\text{общее снижение в RSS, когда выполняются отдельные регрессии} / \text{стоимость в степени свободы}}{\text{общее количество оставшихся RSS выполняются отдельные регрессии} / \text{степеней свободы осталось}}$$

$$= \frac{(RSS_p - [RSS_1 + RSS_2]) / k}{(RSS_1 + RSS_2) / (n - 2k)}$$

Числитель F-статистики состоит из общего улучшения пригонки при расщеплении образца, деленного на стоимость в терминах степеней свободы при выполнении отдельных регрессий.

$$F(k, n - 2k) = \frac{\text{общее снижение в RSS, когда выполняются отдельные регрессии}}{\text{стоимость в степени свободы}} = \frac{\text{общее количество оставшихся RSS выполняются отдельные регрессии}}{\text{степеней свободы осталось}}$$

$$= \frac{(RSS_p - [RSS_1 + RSS_2]) / k}{(RSS_1 + RSS_2) / (n - 2k)}$$

Знаменателем F-статистики является общее количество RSS, оставшееся после расщепления выборки, деленное на количество оставшихся степеней свободы

$$F(k, n - 2k) = \frac{\text{общее снижение в RSS, когда выполняются отдельные регрессии}}{\text{стоимость в степени свободы}} / \frac{\text{общее количество оставшихся RSS выполняются отдельные регрессии}}{\text{степеней свободы осталось}}$$

$$= \frac{(RSS_P - [RSS_1 + RSS_2]) / k}{(RSS_1 + RSS_2) / (n - 2k)}$$

$$RSS_P = 8.91 \times 10^{11}$$

$$RSS_1 + RSS_2 = 4.71 \times 10^{11}$$

$$F(2,70) = \frac{(8.91 \times 10^{11} - [3.49 \times 10^{11} + 1.22 \times 10^{11}]) / 2}{(3.49 \times 10^{11} + 1.22 \times 10^{11}) / 70}$$

В случае функций стоимости школы сокращение остаточной суммы квадратов уже было сведено в таблицу.

$$F(k, n - 2k) = \frac{\text{общее снижение в RSS, когда выполняются отдельные регрессии}}{\text{стоимость в степени свободы}} / \frac{\text{общее количество оставшихся RSS выполняются отдельные регрессии}}{\text{степеней свободы осталось}}$$

$$= \frac{(RSS_P - [RSS_1 + RSS_2]) / k}{(RSS_1 + RSS_2) / (n - 2k)}$$

$$RSS_P = 8.91 \times 10^{11}$$

$$RSS_1 + RSS_2 = 4.71 \times 10^{11}$$

$$F(2, 70) = \frac{(8.91 \times 10^{11} - [3.49 \times 10^{11} + 1.22 \times 10^{11}]) / 2}{(3.49 \times 10^{11} + 1.22 \times 10^{11}) / 70}$$

В модели только два параметра: константа и коэффициент N, поэтому первый аргумент статистики F равен 2

$$F(k, n - 2k) = \frac{\text{общее снижение в RSS, когда выполняются отдельные регрессии}}{\text{стоимость в степени свободы}} / \frac{\text{общее количество оставшихся RSS выполняются отдельные регрессии}}{\text{степеней свободы осталось}}$$

$$= \frac{(RSS_P - [RSS_1 + RSS_2]) / k}{(RSS_1 + RSS_2) / (n - 2k)}$$

$$RSS_P = 8.91 \times 10^{11}$$

$$RSS_1 + RSS_2 = 4.71 \times 10^{11}$$

$$F(2,70) = \frac{(8.91 \times 10^{11} - [3.49 \times 10^{11} + 1.22 \times 10^{11}]) / 2}{(3.49 \times 10^{11} + 1.22 \times 10^{11}) / 70}$$

Остаточная сумма квадратов, оставшихся после расщепления образца, представляет собой сумму RSS1 и RSS2.

$$F(k, n - 2k) = \frac{\text{общее снижение в RSS, когда выполняются отдельные регрессии}}{\text{стоимость в степени свободы}} / \frac{\text{общее количество оставшихся RSS выполняются отдельные регрессии}}{\text{степеней свободы осталось}}$$

$$= \frac{(RSS_P - [RSS_1 + RSS_2]) / k}{(RSS_1 + RSS_2) / (n - 2k)}$$

$$RSS_P = 8.91 \times 10^{11}$$

$$RSS_1 + RSS_2 = 4.71 \times 10^{11}$$

$$F(2,70) = \frac{(8.91 \times 10^{11} - [3.49 \times 10^{11} + 1.22 \times 10^{11}]) / 2}{(3.49 \times 10^{11} + 1.22 \times 10^{11}) / 70} = 31.2$$

$$F(2,70)_{\text{crit}, 0.1\%} = 7.6$$

Таким образом, статистика F равна 31.2. Критическое значение F (2,70) составляет 7,6 при уровне значимости 0,1%.

$$F(k, n - 2k) = \frac{\begin{array}{l} \text{общее снижение в RSS, когда} \\ \text{выполняются отдельные регрессии} \end{array}}{\begin{array}{l} \text{общее количество оставшихся RSS} \\ \text{выполняются отдельные регрессии} \end{array}} \bigg/ \frac{\begin{array}{l} \text{стоимость в степени} \\ \text{свободы} \end{array}}{\begin{array}{l} \text{степеней свободы} \\ \text{осталось} \end{array}}$$

$$= \frac{(RSS_P - [RSS_1 + RSS_2]) / k}{(RSS_1 + RSS_2) / (n - 2k)}$$

$$\begin{aligned} RSS_P &= 8.91 \times 10^{11} \\ RSS_1 + RSS_2 &= 4.71 \times 10^{11} \end{aligned}$$

$$F(2,70) = \frac{(8.91 \times 10^{11} - [3.49 \times 10^{11} + 1.22 \times 10^{11}]) / 2}{(3.49 \times 10^{11} + 1.22 \times 10^{11}) / 70} = 31.2$$

$$F(2,70)_{\text{crit}, 0.1\%} = 7.6$$

Поэтому уменьшение остаточной суммы квадратов является значительным на уровне 0,1%. Мы пришли к выводу, что функция объединенных затрат является неадекватной спецификацией и что мы должны проводить отдельные регрессии для двух типов школ