МУЛЬТИКОЛЛИНЕАРНОСТЬ

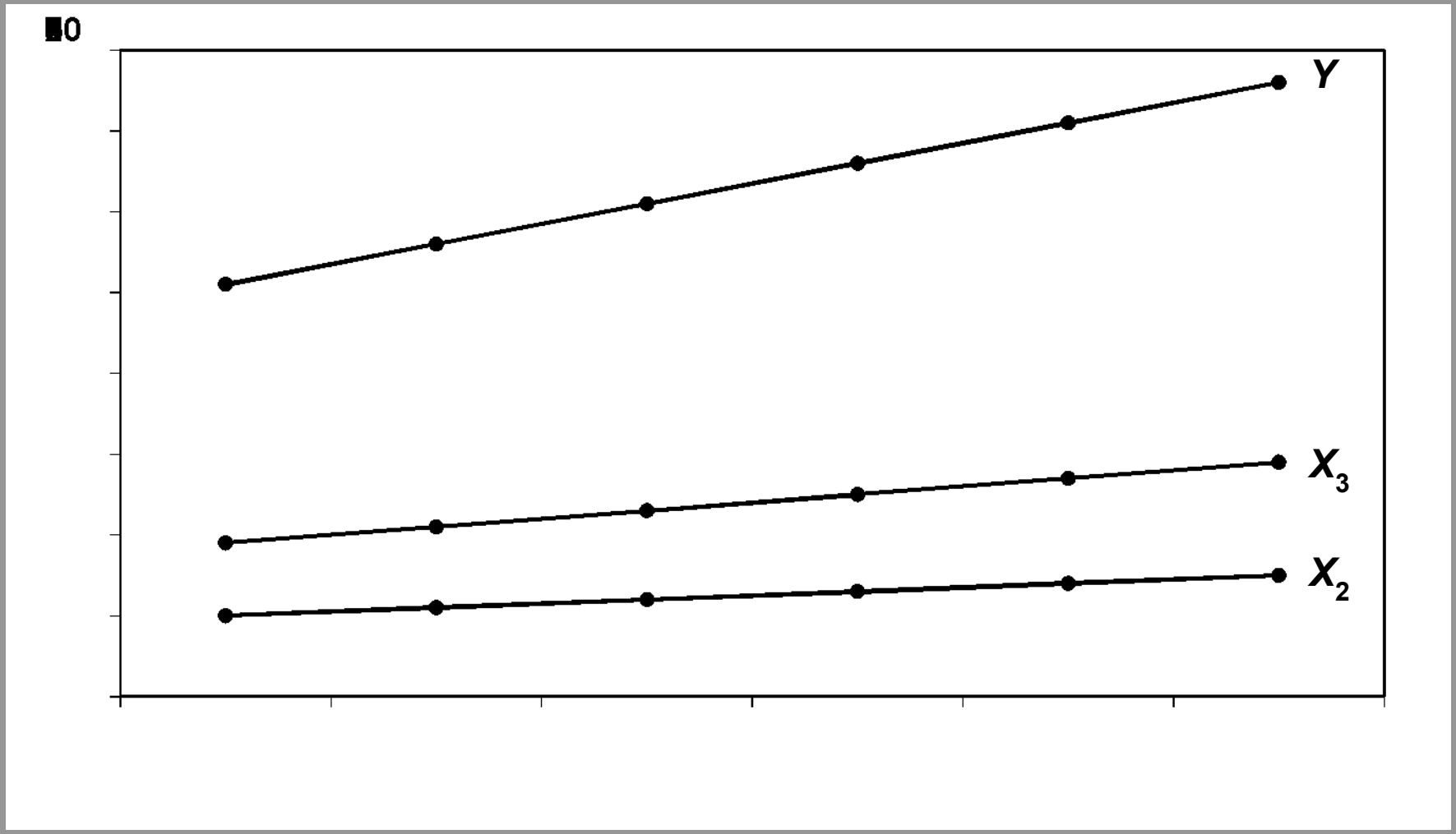
$$Y = 2 + 3X_2 + X_3$$

$$X_3 = 2X_2 - 1$$

X_2	X_3	Y
10	19	51
11	21	56
12	23	61
13	25	66
14	27	71
15	29	76

Предположим, что Y = 2 + 3X2 + X3 и что X3 = 2X2 - 1. В уравнении для Y нет срока нарушения, но это не важно. Предположим, что у нас есть шесть приведенных наблюдений.

МУЛЬТИКОЛЛИНЕАРНОСТЬ



Три переменные отображаются в виде линейных графиков выше. Рассматривая данные, невозможно определить, вызваны ли изменения в Y изменениями X2, изменениями в X3 или совместно изменениями X2 и X3.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

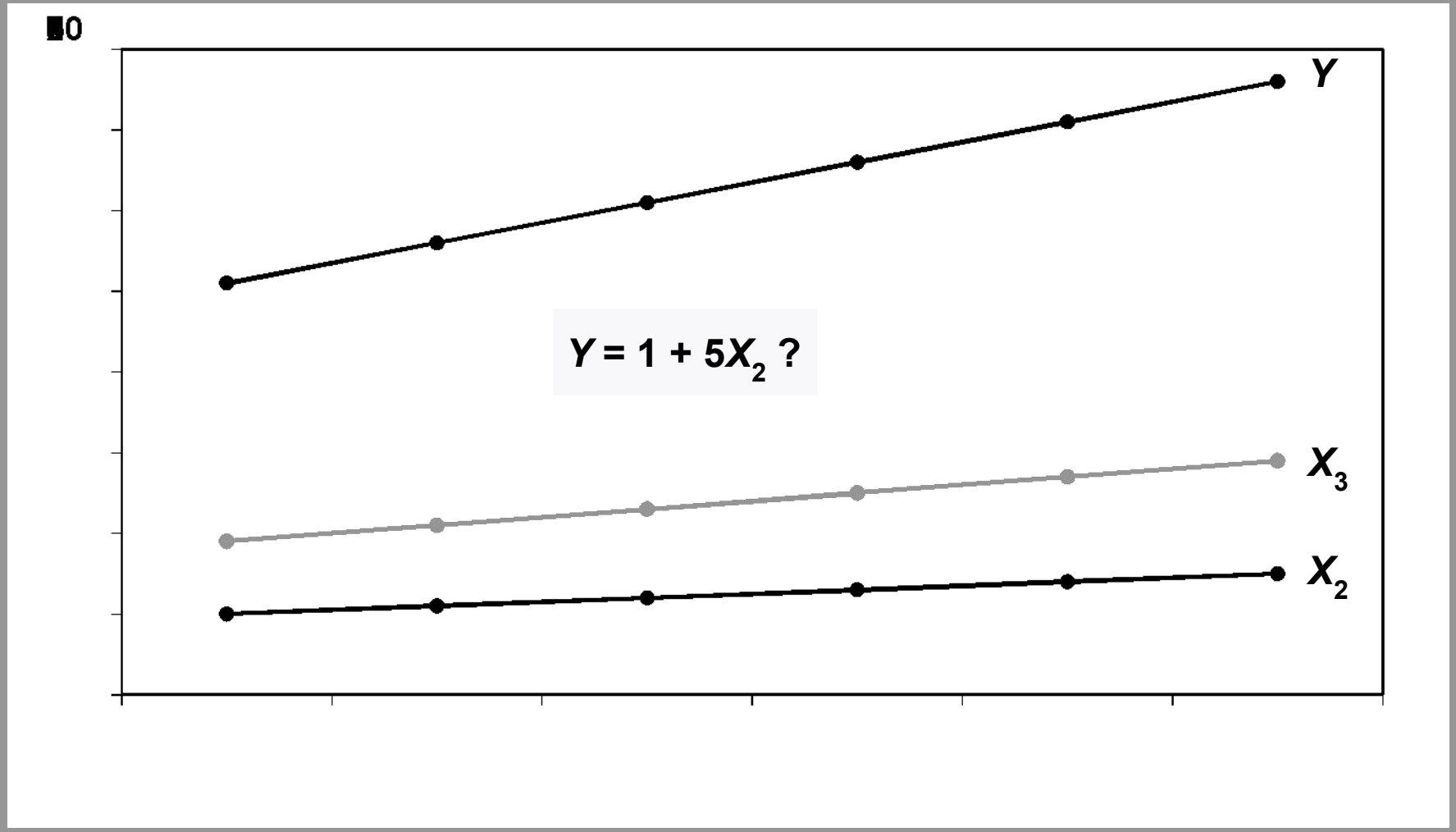
$$Y = 2 + 3X_2 + X_3$$

$$X_3 = 2X_2 - 1$$

X_2	X_3	Y	X_2	X_3	Y
			изменение от предыдущего наблюдения		
10	19	51			
11	21	56	1	2	5
12	23	61	1	2	5
13	25	66	1	2	5
14	27	71	1	2	5
15	29	76	1	2	5

Численно Y увеличивается на 5 в каждом наблюдении. X2 изменяется на 1.

МУЛЬТИКОЛЛИНЕАРНОСТЬ



Следовательно, истинное соотношение могло бы быть Y = 1 + 5X2.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

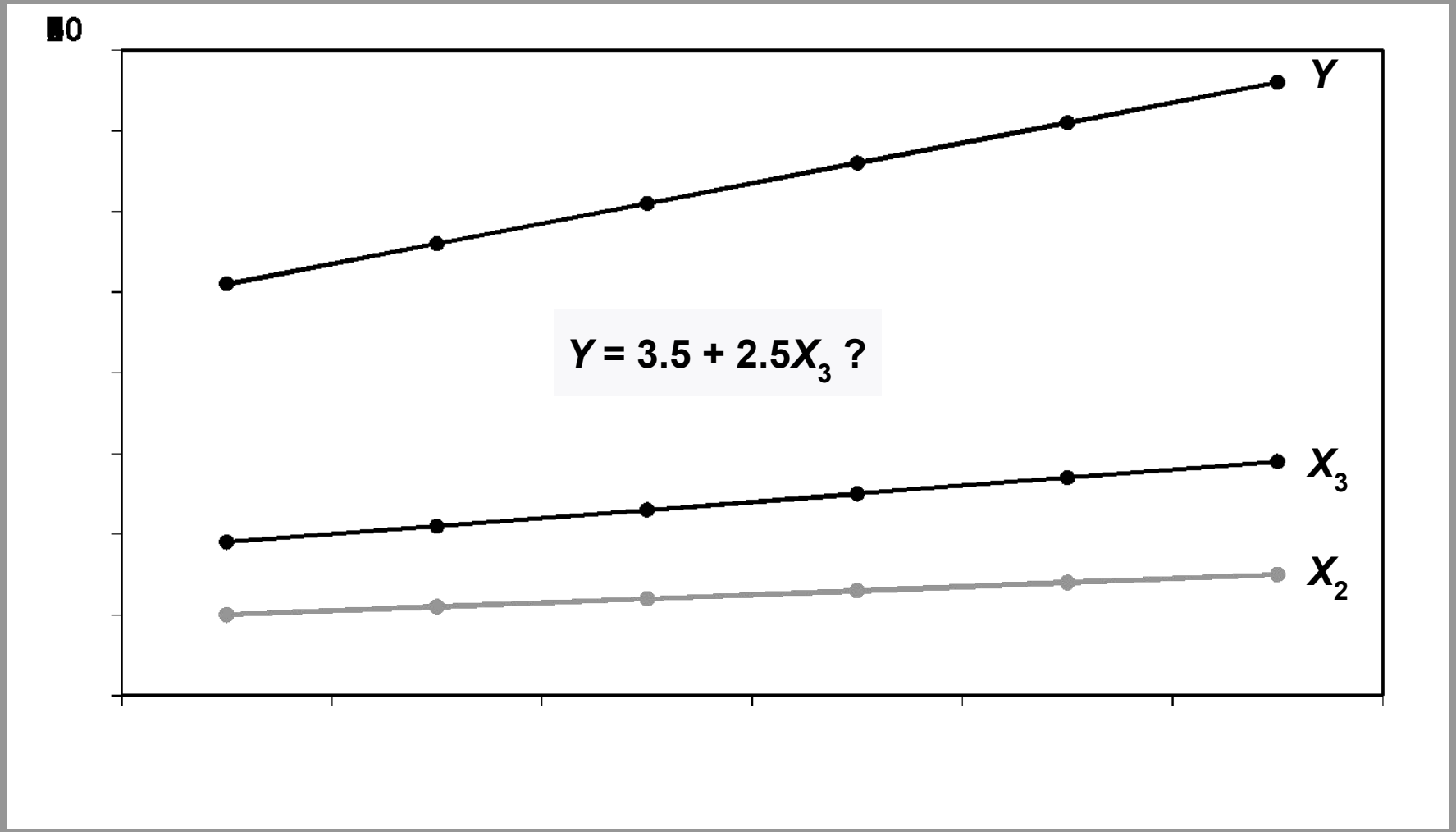
$$Y = 2 + 3X_2 + X_3$$

$$X_3 = 2X_2 - 1$$

X_2	X_3	Y						X_2	X_3	Y
			изменение от предыдущего наблюдения							
10	19	51								
11	21	56	1	2	5					
12	23	61	1	2	5					
13	25	66	1	2	5					
14	27	71	1	2	5					
15	29	76	1	2	5					

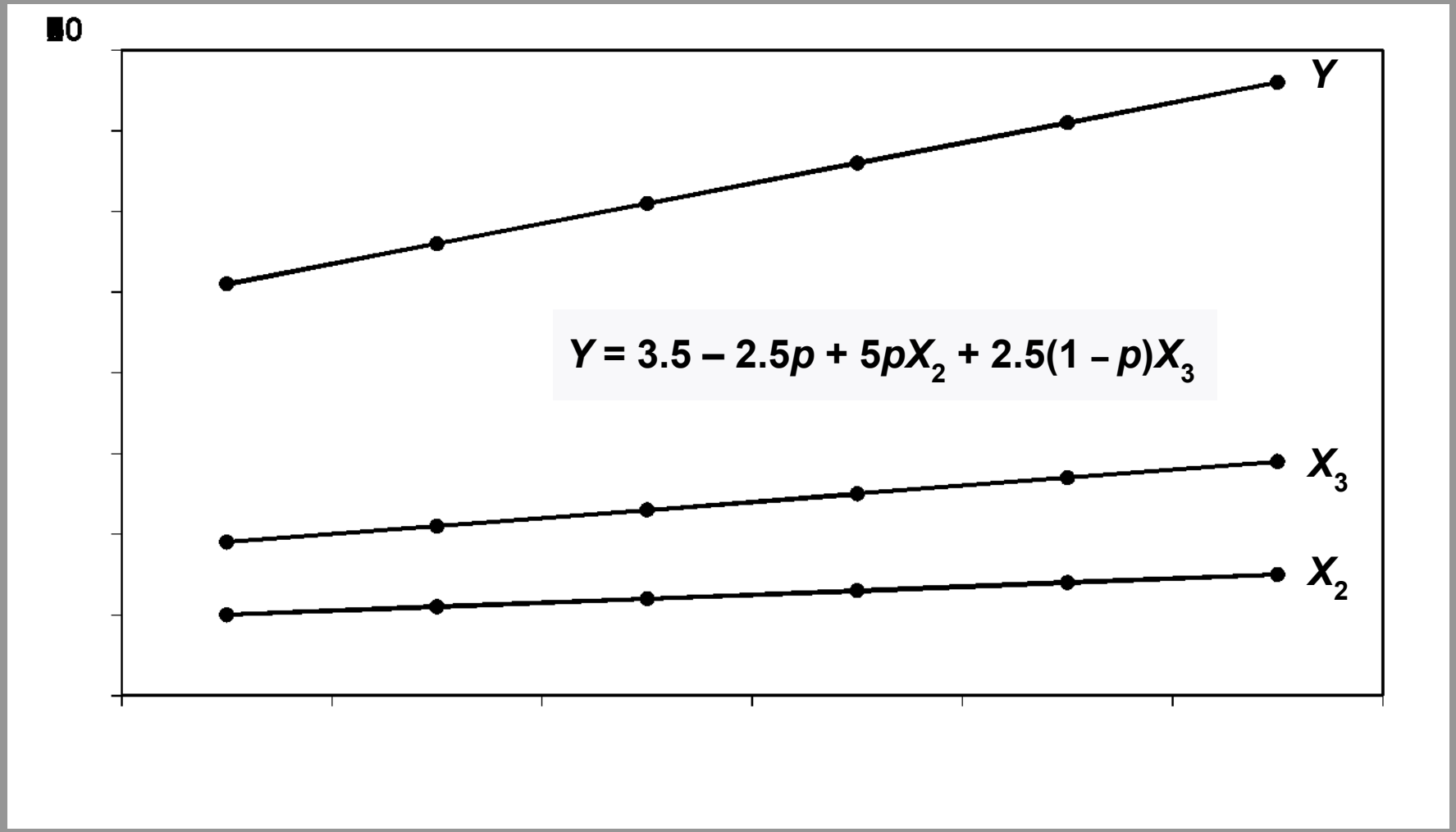
Однако также можно видеть, что X3 увеличивается на 2 в каждом наблюдении.

МУЛЬТИКОЛЛИНЕАРНОСТЬ



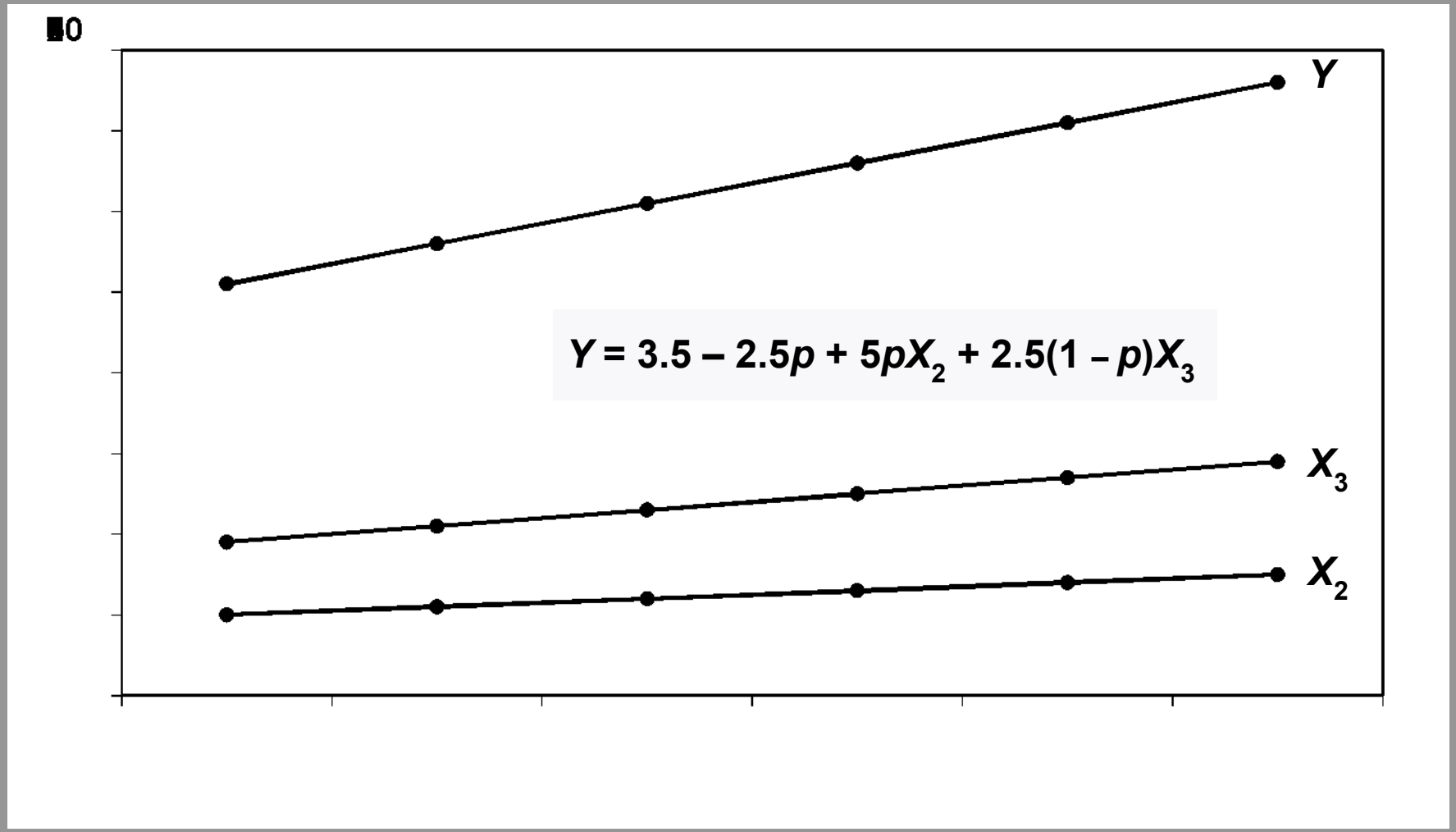
Следовательно, истинная связь могла бы быть Y = 3,5 + 2,5X3.

МУЛЬТИКОЛЛИНЕАРНОСТЬ



Эти две возможности являются частными случаями Y = 3,5 - 2,5p + 5pX2 + 2,5 (1 - p) X3, которые соответствовали бы соотношению для любого значения p.

МУЛЬТИКОЛЛИНЕАРНОСТЬ



Нет никакого способа, чтобы регрессионный анализ или любая другая техника могли определять истинную связь из этого бесконечного множества возможностей, учитывая данные выборки.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

$$X_3 = \lambda + \mu X_2$$

Что произойдет, если вы попытаетесь запустить регрессию, когда существует точная линейная зависимость между объясняющими переменными?

МУЛЬТИКОЛЛИНЕАРНОСТЬ

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad X_3 = \lambda + \mu X_2$$

Мы исследуем, используя модель с двумя объясняющими переменными, показанными выше. [Примечание: термин «нарушение» теперь включен в истинную модель, но это не имеет никакого значения для анализа.]

МУЛЬТИКОЛЛИНЕАРНОСТЬ

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad X_3 = \lambda + \mu X_2$$

$$\hat{\beta}_2 = \frac{\sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y}) \sum (X_{3i} - \bar{X}_3)^2 - \sum (X_{3i} - \bar{X}_3)(Y_i - \bar{Y}) \sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2 \sum (X_{3i} - \bar{X}_3)^2 - \left(\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)\right)^2}$$

Выражение для коэффициента множественной регрессии b2 показано выше. Мы заменим X3, используя его связь с X2.

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad X_3 = \lambda + \mu X_2$$

$$\hat{\beta}_2 = \frac{\sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y}) \sum (X_{3i} - \bar{X}_3)^2 - \sum (X_{3i} - \bar{X}_3)(Y_i - \bar{Y}) \sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2 \sum (X_{3i} - \bar{X}_3)^2 - \left(\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)\right)^2}$$

$$\sum (X_{3i} - \bar{X}_3)^2 = \sum \left([\lambda + \mu X_{2i}] - [\lambda + \mu \bar{X}_2]\right)^2$$

$$= \sum (\mu X_{2i} - \mu \bar{X}_2)^2 = \sum \mu^2 (X_{2i} - \bar{X}_2)^2$$

$$= \mu^2 \sum (X_{2i} - \bar{X}_2)^2$$

Во-первых, мы заменим термины, выделенные.

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad X_3 = \lambda + \mu X_2$$

$$\hat{\beta}_2 = \frac{\sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})\mu^2 \sum (X_{2i} - \bar{X}_2)^2 - \sum (X_{3i} - \bar{X}_3)(Y_i - \bar{Y}) \sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2 \mu^2 \sum (X_{2i} - \bar{X}_2)^2 - \left(\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)\right)^2}$$

$$\sum (X_{3i} - \bar{X}_3)^2 = \sum ([\lambda + \mu X_{2i}] - [\lambda + \mu \bar{X}_2])^2$$

$$= \sum (\mu X_{2i} - \mu \bar{X}_2)^2 = \sum \mu^2 (X_{2i} - \bar{X}_2)^2$$

$$= \mu^2 \sum (X_{2i} - \bar{X}_2)^2$$

Мы сделали замену.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad X_3 = \lambda + \mu X_2$$

$$\hat{\beta}_2 = \frac{\sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})\mu^2 \sum (X_{2i} - \bar{X}_2)^2 - \sum (X_{3i} - \bar{X}_3)(Y_i - \bar{Y}) \sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2 \mu^2 \sum (X_{2i} - \bar{X}_2)^2 - \left(\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)\right)^2}$$

$$\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3) = \sum (X_{2i} - \bar{X}_2)([\lambda + \mu X_{2i}] - [\lambda + \mu \bar{X}_2])$$

$$= \sum (X_{2i} - \bar{X}_2)(\mu X_{2i} - \mu \bar{X}_2)$$

$$= \mu \sum (X_{2i} - \bar{X}_2)^2$$

Далее, термины, выделенные сейчас.

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad X_3 = \lambda + \mu X_2$$

$$\hat{\beta}_2 = \frac{\sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})\mu^2 \sum (X_{2i} - \bar{X}_2)^2 - \sum (X_{3i} - \bar{X}_3)(Y_i - \bar{Y})\mu \sum (X_{2i} - \bar{X}_2)^2}{\sum (X_{2i} - \bar{X}_2)^2 \mu^2 \sum (X_{2i} - \bar{X}_2)^2 - \left(\mu \sum (X_{2i} - \bar{X}_2)^2\right)^2}$$

$$\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3) = \sum (X_{2i} - \bar{X}_2)([\lambda + \mu X_{2i}] - [\lambda + \mu \bar{X}_2])$$

$$= \sum (X_{2i} - \bar{X}_2)(\mu X_{2i} - \mu \bar{X}_2)$$

$$= \mu \sum (X_{2i} - \bar{X}_2)^2$$

Мы сделали замену.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad X_3 = \lambda + \mu X_2$$

$$\hat{\beta}_2 = \frac{\sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})\mu^2 \sum (X_{2i} - \bar{X}_2)^2 - \sum (X_{3i} - \bar{X}_3)(Y_i - \bar{Y})\mu \sum (X_{2i} - \bar{X}_2)^2}{\sum (X_{2i} - \bar{X}_2)^2 \mu^2 \sum (X_{2i} - \bar{X}_2)^2 - \left(\mu \sum (X_{2i} - \bar{X}_2)^2\right)^2}$$

$$\begin{aligned}\sum (X_{3i} - \bar{X}_3)(Y_i - \bar{Y}) &= \sum ([\lambda + \mu X_{2i}] - [\lambda + \mu \bar{X}_2])(Y_i - \bar{Y}) \\ &= \sum (\mu X_{2i} - \mu \bar{X}_2)(Y_i - \bar{Y}) \\ &= \mu \sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})\end{aligned}$$

Наконец, этот термин.

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad X_3 = \lambda + \mu X_2$$

$$\hat{\beta}_2 = \frac{\sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})\mu^2 \sum (X_{2i} - \bar{X}_2)^2 - \mu \sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y}) \mu \sum (X_{2i} - \bar{X}_2)^2}{\sum (X_{2i} - \bar{X}_2)^2 \mu^2 \sum (X_{2i} - \bar{X}_2)^2 - \left(\mu \sum (X_{2i} - \bar{X}_2)^2 \right)^2}$$

$$\begin{aligned} \sum (X_{3i} - \bar{X}_3)(Y_i - \bar{Y}) &= \sum ([\lambda + \mu X_{2i}] - [\lambda + \mu \bar{X}_2])(Y_i - \bar{Y}) \\ &= \sum (\mu X_{2i} - \mu \bar{X}_2)(Y_i - \bar{Y}) \\ &= \mu \sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y}) \end{aligned}$$

Опять же, мы сделали замену.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad X_3 = \lambda + \mu X_2$$

$$\hat{\beta}_2 = \frac{\sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})\mu^2 \sum (X_{2i} - \bar{X}_2)^2 - \mu \sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})\mu \sum (X_{2i} - \bar{X}_2)^2}{\sum (X_{2i} - \bar{X}_2)^2 \mu^2 \sum (X_{2i} - \bar{X}_2)^2 - \left(\mu \sum (X_{2i} - \bar{X}_2)^2 \right)^2}$$

$$= \frac{0}{0}$$

Оказывается, что числитель и знаменатель равны нулю. Коэффициент регрессии не определен.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad X_3 = \lambda + \mu X_2$$

$$\hat{\beta}_2 = \frac{\sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})\mu^2 \sum (X_{2i} - \bar{X}_2)^2 - \mu \sum (X_{2i} - \bar{X}_2)(Y_i - \bar{Y})\mu \sum (X_{2i} - \bar{X}_2)^2}{\sum (X_{2i} - \bar{X}_2)^2 \mu^2 \sum (X_{2i} - \bar{X}_2)^2 - \left(\mu \sum (X_{2i} - \bar{X}_2)^2 \right)^2}$$

$$= \frac{0}{0}$$

Необычно, что существует точная взаимосвязь между объясняющими переменными в регрессии. Когда это происходит, это типично, потому что в спецификации есть логическая ошибка.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

```
. reg EARNINGS S EXP
```

Source	SS	df	MS	Number of obs	500
Model	8735.42401	2	4367.712	F(2, 497)	35.24
Residual	61593.5422	497	123.930668	Prob > F	0.0000
Total	70328.9662	499	140.939812	R-squared	0.1242
				Adj R-squared	0.1207
				Root MSE	11.132

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	1.877563	.2237434	8.39	0.000	1.437964	2.317163
EXP	.9833436	.2098457	4.69	0.000	.5710495	1.395638
_cons	-14.66833	4.288375	-3.42	0.001	-23.09391	-6.242752

Однако часто бывает, что существует приблизительная взаимосвязь. Мы будем использовать уравнение заработной платы в качестве иллюстрации.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

```
. reg EARNINGS S EXP
```

Source	SS	df	MS	Number of obs	500
Model	8735.42401	2	4367.712	F(2, 497)	35.24
Residual	61593.5422	497	123.930668	Prob > F	0.0000
Total	70328.9662	499	140.939812	R-squared	0.1242
				Adj R-squared	0.1207
				Root MSE	11.132

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	1.877563	.2237434	8.39	0.000	1.437964	2.317163
EXP	.9833436	.2098457	4.69	0.000	.5710495	1.395638
_cons	-14.66833	4.288375	-3.42	0.001	-23.09391	-6.242752

Когда вы связываете заработную плату с учебой и опытом работы, она, если часто разумно предположить, что влияние опыта работы подлежит уменьшению.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

```
. reg EARNINGS S EXP
```

```
------------------------------------------------------------------------------
      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =   35.24
       Model |  8735.42401     2    4367.712           Prob > F      =  0.0000
    Residual |  61593.5422   497  123.930668           R-squared     =  0.1242
-------------+------------------------------           Adj R-squared =  0.1207
       Total |  70328.9662   499  140.939812           Root MSE      =  11.132
------------------------------------------------------------------------------
    EARNINGS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           S |   1.877563   .2237434     8.39   0.000     1.437964    2.317163
         EXP |   .9833436   .2098457     4.69   0.000     .5710495    1.395638
       _cons |  -14.66833   4.288375    -3.42   0.001    -23.09391   -6.242752
------------------------------------------------------------------------------
```

Стандартный способ разрешить это - включить EXPSQ, квадрат EXP, в спецификацию. Согласно гипотезе о снижении доходности, коэффициент EXPSQ должен быть отрицательным.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

```
. reg EARNINGS S EXP EXPSQ
```

Source	SS	df	MS	Number of obs	500
Model	8793.741	3	2931.247	F(3, 496)	23.63
Residual	61535.2252	496	124.062954	Prob > F	0.0000
				R-squared	0.1250
				Adj R-squared	0.1197
Total	70328.9662	499	140.939812	Root MSE	11.138

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	1.869284	.2241882	8.34	0.000	1.428809	2.30976
EXP	1.427853	.6814907	2.10	0.037	.0888882	2.766817
EXPSQ	-.0328379	.047896	-0.69	0.493	-.126942	.0612662
_cons	-15.7658	4.57953	-3.44	0.001	-24.76347	-6.76813

Мы вписываем эту спецификацию с помощью набора данных 21

МУЛЬТИКОЛЛИНЕАРНОСТЬ

```
. reg EARNINGS S EXP EXPSQ
```

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	1.869284	.2241882	8.34	0.000	1.428809	2.30976
EXP	1.427853	.6814907	2.10	0.037	.0888882	2.766817
EXPSQ	-.0328379	.047896	-0.69	0.493	-.126942	.0612662
_cons	-15.7658	4.57953	-3.44	0.001	-24.76347	-6.76813

```
. reg EARNINGS S EXP
```

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	1.877563	.2237434	8.39	0.000	1.437964	2.317163
EXP	.9833436	.2098457	4.69	0.000	.5710495	1.395638
_cons	-14.66833	4.288375	-3.42	0.001	-23.09391	-6.242752

Школьный компонент результатов регрессии мало влияет на включение термина EXPSQ. Коэффициент S указывает, что дополнительный год обучения увеличивается почасовой заработок на 1,88 доллара. В спецификации без EXPSQ было 1,87.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

```
. reg EARNINGS S EXP EXPSQ
```

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	1.869284	.2241882	8.34	0.000	1.428809	2.30976
EXP	1.427853	.6814907	2.10	0.037	.0888882	2.766817
EXPSQ	-.0328379	.047896	-0.69	0.493	-.126942	.0612662
_cons	-15.7658	4.57953	-3.44	0.001	-24.76347	-6.76813

```
. reg EARNINGS S EXP
```

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	1.877563	.2237434	8.39	0.000	1.437964	2.317163
EXP	.9833436	.2098457	4.69	0.000	.5710495	1.395638
_cons	-14.66833	4.288375	-3.42	0.001	-23.09391	-6.242752

Аналогично, стандартная ошибка 0.22 в спецификации без EXPSQ также мало изменилась, и коэффициент остается очень значительным.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

```
. reg EARNINGS S EXP EXPSQ
```

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	1.869284	.2241882	8.34	0.000	1.428809	2.30976
EXP	1.427853	.6814907	2.10	0.037	.0888882	2.766817
EXPSQ	-.0328379	.047896	-0.69	0.493	-.126942	.0612662
_cons	-15.7658	4.57953	-3.44	0.001	-24.76347	-6.76813

```
. reg EARNINGS S EXP
```

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	1.877563	.2237434	8.39	0.000	1.437964	2.317163
EXP	.9833436	.2098457	4.69	0.000	.5710495	1.395638
_cons	-14.66833	4.288375	-3.42	0.001	-23.09391	-6.242752

В спецификации без EXPSQ коэффициент EXP значителен на уровне 0,1 процента. Когда добавляется EXPSQ, это значимо только на уровне 5 процентов.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

```
. reg EARNINGS S EXP EXPSQ
```

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	1.869284	.2241882	8.34	0.000	1.428809	2.30976
EXP	1.427853	.6814907	2.10	0.037	.0888882	2.766817
EXPSQ	-.0328379	.047896	-0.69	0.493	-.126942	.0612662
_cons	-15.7658	4.57953	-3.44	0.001	-24.76347	-6.76813

```
. reg EARNINGS S EXP
```

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	1.877563	.2237434	8.39	0.000	1.437964	2.317163
EXP	.9833436	.2098457	4.69	0.000	.5710495	1.395638
_cons	-14.66833	4.288375	-3.42	0.001	-23.09391	-6.242752

Это связано главным образом с тем, что стандартная ошибка увеличилась с 0,21 до 0,68, что свидетельствует о значительной потере точности.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

```
. reg EARNINGS S EXP EXPSQ
```

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	1.869284	.2241882	8.34	0.000	1.428809	2.30976
EXP	1.427853	.6814907	2.10	0.037	.0888882	2.766817
EXPSQ	-.0328379	.047896	-0.69	0.493	-.126942	.0612662
_cons	-15.7658	4.57953	-3.44	0.001	-24.76347	-6.76813

```
. reg EARNINGS S EXP
```

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	1.877563	.2237434	8.39	0.000	1.437964	2.317163
EXP	.9833436	.2098457	4.69	0.000	.5710495	1.395638
_cons	-14.66833	4.288375	-3.42	0.001	-23.09391	-6.242752

В исходной спецификации 95-процентный доверительный интервал для коэффициента EXP составлял от 0,57 до 1,40, что уже достаточно свободно. Теперь это от 0.09 до 2.76.

МУЛЬТИКОЛЛИНЕАРНОСТЬ

```
. reg EARNINGS S EXP EXPSQ
```

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	1.869284	.2241882	8.34	0.000	1.428809	2.30976
EXP	1.427853	.6814907	2.10	0.0		
EXPSQ	-.0328379	.047896	-0.69	0.4		
_cons	-15.7658	4.57953	-3.44	0.0		

```
. cor EXP EXPSQ
(obs=500)
```

	EXP	EXPSQ
EXP	1.0000	
EXPSQ	0.9677	1.0000

```
. reg EARNINGS S EXP
```

EARNINGS	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
S	1.877563	.2237434	8.39	0.000	1.437964	2.317163
EXP	.9833436	.2098457	4.69	0.000	.5710495	1.395638
_cons	-14.66833	4.288375	-3.42	0.001	-23.09391	-6.242752

Потеря точности связана с мультиколлинеарностью, корреляция между EXP и EXPSQ составляет 0,97. Коэффициент EXPSQ имеет ожидаемый отрицательный знак, но он не является отдаленно значимым.