

МОДЕЛЬ ПАРНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

№ домохозяйства	Среднедушевой доход домохозяйства, д.е.	Объем спроса, ед.
1	100	24
2	200	42
3	150	35
4	80	24
5	160	39

Сервис – Анализ данных

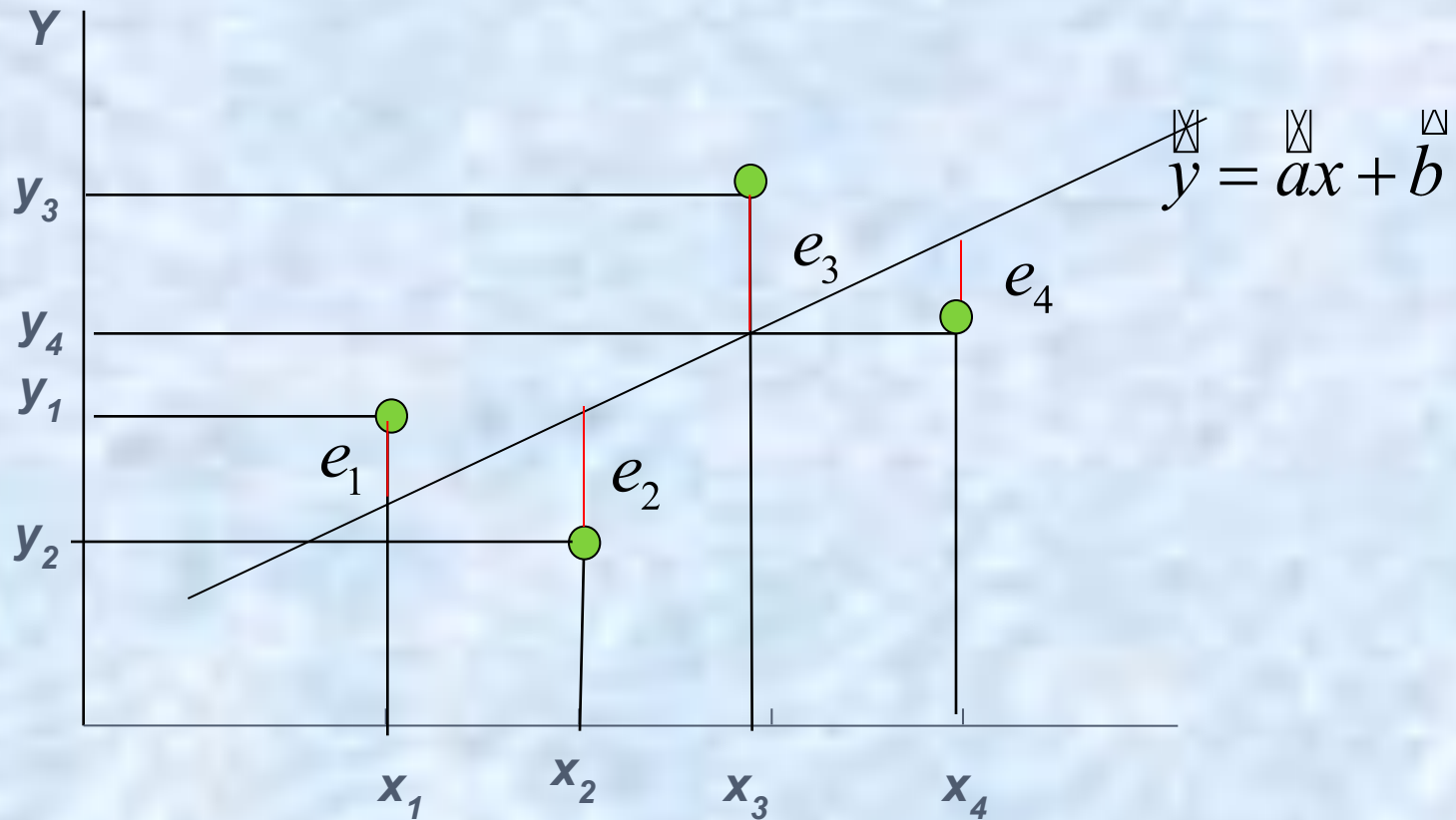
	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение
Y-пересечение	9,334052	3,296116	2,831833	0,06609
Переменная X 1	0,170043	0,0228	7,458124	0,004991

$y=0,17x+9,33$ - функция спроса в зависимости от дохода.

С ростом дохода на 1 ден.ед. спрос на товар растет на 0,17 ед.

ПОКАЗАТЕЛИ КАЧЕСТВА УРАВНЕНИЯ РЕГРЕССИИ

Все показатели качества используют остатки:



$$e_i = y_i - \hat{y}_i = y_i - \hat{a}x_i - \hat{b} \text{ называется } \underline{\text{остатком}}$$

Как оценить качество построенной модели?

Построим прогноз по модели по формуле $\hat{y} = 0,17x + 9,33$

	А	В	С
1	X	y	прогноз $y = 0,17x + 9,3$
2	100	24	26,338
3	200	42	43,343
4	150	35	34,841
5	80	24	22,938
6	160	39	36,541

Как оценить качество построенной модели?

Вычисляем остатки $e = y - \hat{y}$

	A	B	C	D
	X	y	прогноз $y=0,17x+9,3$	e
1	100	24	26,338	-2,338
2	200	42	43,343	-1,343
3	150	35	34,841	0,1595
4	80	24	22,938	1,0625
5	160	39	36,541	2,4591

Как оценить качество построенной модели?

Находим относительную ошибку аппроксимации

$$A = \frac{|y - \hat{y}|}{y}$$

	A	B	C	D	E
	X	y	прогноз $y=0,17x+9,3$	e	Относительная ошибка
1					
2	100	24	26,338	-2,338	9,74%
3	200	42	43,343	-1,343	3,20%
4	150	35	34,841	0,1595	0,46%
5	80	24	22,938	1,0625	4,43%
6	160	39	36,541	2,4591	6,31%

Процентный формат

Как оценить качество построенной модели?

Находим среднюю относительную ошибку аппроксимации

	A	B	C	D	E
1	X	y	прогноз $y=0,17x+9,3$	e	Относительная ошибка
2	100	24	26,338	-2,338	9,74%
3	200	42	43,343	-1,343	3,20%
4	150	35	34,841	0,1595	0,46%
5	80	24	22,938	1,0625	4,43%
6	160	39	36,541	2,4591	6,31%
7			Средняя относительная ошибка		4,83%
8					

↑
среднее по столбцу

В среднем прогноз отличается от наблюдаемого значения на 4,83%

Как оценить качество построенной модели?

Еще один показатель качества – коэффициент детерминации
Для его вычисления вычисляем сумму квадратов остатков ESS
(Error Sum of Squares)

	A	B	C	D	E	F
1	X	y	прогноз $y=0,17x+9,3$	e	Относительная ошибка	e^2
2	100	24	26,338	-2,338	9,74%	5,468
3	200	42	43,343	-1,343	3,20%	1,803
4	150	35	34,841	0,1595	0,46%	0,025
5	80	24	22,938	1,0625	4,43%	1,129
6	160	39	36,541	2,4591	6,31%	6,047
7			Средняя относительная ошибка		4,83%	14,472

↑
Сумма по столбцу

Вычисляем суммы квадратов

- остаточную, регрессионную и общую

$$ESS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \text{— error sum of squares, остаточная сумма квадратов отклонений}$$

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2 = ns_y^2 \quad \text{— total sum of squares, общая сумма квадратов отклонений}$$

$$RSS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad \text{— regression sum of squares, регрессионная сумма квадратов отклонений}$$

Теорема $TSS = RSS + ESS$

Коэффициент детерминации R^2

Коэффициентом детерминации называют число

$$R^2 = \frac{RSS}{TSS}$$

$$R^2 = \frac{TSS - ESS}{TSS} = 1 - \frac{ESS}{TSS}$$

Коэффициентом детерминации - это доля дисперсии признака y , объясненная регрессией в общей дисперсии признака y

Свойства коэффициента детерминации R^2

$$R^2 = \frac{RSS}{TSS}$$

1) $0 \leq R^2 \leq 1$

Свойства коэффициента детерминации R^2

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

2) Если $R^2 = 1$, то линия регрессии точно соответствует всем наблюдениям и между x и y существует линейная функциональная зависимость.

Свойства коэффициента детерминации R^2

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

3) Если $R^2 = 0$, то линия регрессии параллельна оси абсцисс и переменная y не зависит от переменной x .

$\bar{y}$

Свойства коэффициента детерминации R^2

$$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

4) В случае парной линейной регрессии $R^2 = (\text{cor}(x, y))^2$

	A	B	C	D	E	F
1	X	y	прогноз y=0,17x+9,3	e	Относитель ная ошибка	e^2
2	100	24	26,338	-2,338	9,74%	5,468
3	200	42	43,343	-1,343	3,20%	1,803
4	150	35	34,841	0,1595	0,46%	0,025
5	80	24	22,938	1,0625	4,43%	1,129
6	160	39	36,541	2,4591	6,31%	6,047
7			Средняя относительная ошибка		4,83%	14,472

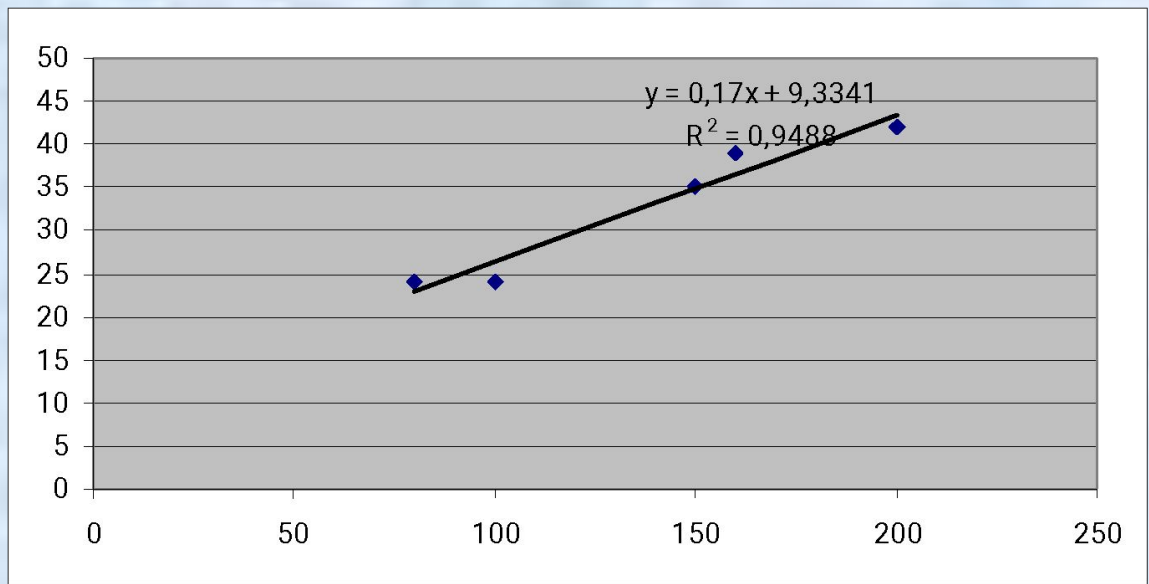
ДИСПР(y)

$$R^2 = 1 - \frac{ESS}{n \cdot s_y^2}$$

R2=	0,949
-----	-------

94,9% вариации спроса на продукт объясняется доходом и остальные 5,1% прочими факторами, не включенными в модель

R2=		0,949	



Стандартная ошибка регрессии (средняя квадратическая ошибка)

s

$$s = \sqrt{\frac{1}{n-2} ESS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n \left(y_i - \bar{a}x_i - \bar{b} \right)^2}$$

Пример построения уравнения регрессии

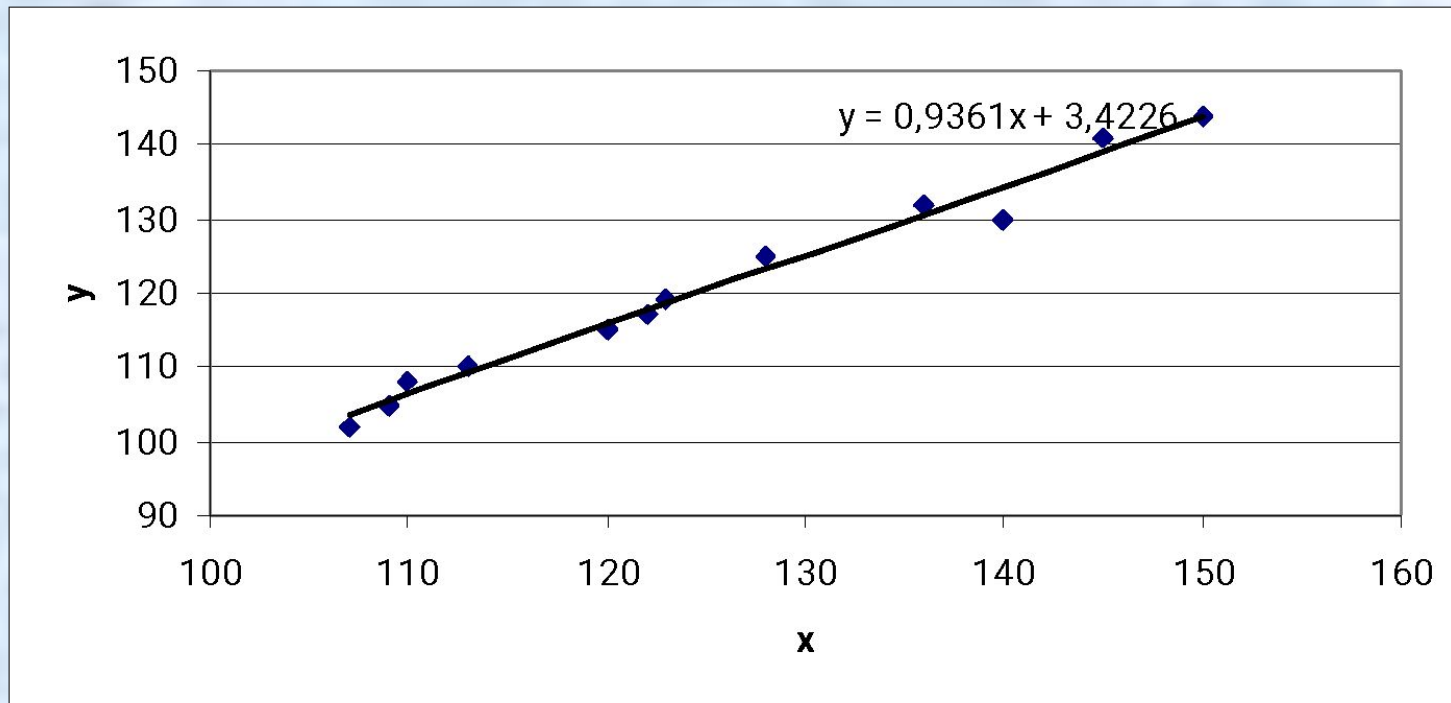
При анализе зависимости объема потребления Y (у.е.) домохозяйства от располагаемого дохода X (у.е.) отобрана выборка объема $n = 12$ (помесячно в течение года), результаты которой приведены в таблице:

i	1	2	3	4	5	6	7	8	9	10	11	12
x_i	107	109	110	113	120	122	123	128	136	140	145	150
y_i	102	105	108	110	115	117	119	125	132	130	141	144

Пример построения уравнения регрессии

Т.о., уравнение парной линейной регрессии имеет вид:

$$\hat{Y} = 0,9361X + 3,423$$



Пример построения уравнения регрессии

Вывод итогов	
Регрессионная статистика	
Множеств	0,991607
R-квадрат	0,983284
Нормиров	0,981612
Стандартн	1,877467
Наблюден	12

$R^2 = 0,98$

$s = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - ax_i - b)^2}$

Дисперсионный анализ					
	df	SS	MS	F	значимость F
Регрессия	1	2073,418	2073,418	588,2235	3,23E-10
Остаток	10	35,24881	3,524881		
Итого	11	2108,667			

RSS

ESS

TSS

Коэффициенты регрессии, стандартная ошибка, статистика t, p-значения, нижние 95% и верхние 95%						
Y-пересеч	3,42261	4,864432	0,703599	0,497738	-7,41602	14,26124
x	0,93608	0,038596	24,25332	3,23E-10	0,850083	1,022077

$$\hat{Y} = 0,9361X + 3,423$$

x	y	y прогноз	e	A
107	102	103,5832	-1,5832	1,55%
109	105	105,4554	-0,45536	0,43%
110	108	106,3914	1,608558	1,49%
113	110	109,1997	0,800317	0,73%
120	115	115,7522	-0,75225	0,65%
122	117	117,6244	-0,62441	0,53%
123	119	118,5605	0,439514	0,37%
128	125	123,2409	1,759113	1,41%
136	132	130,7295	1,27047	0,96%
140	130	134,4739	-4,47385	3,44%
145	141	139,1543	1,845747	1,31%
150	144	143,8347	0,165346	0,11%
			Среднее	1,08%

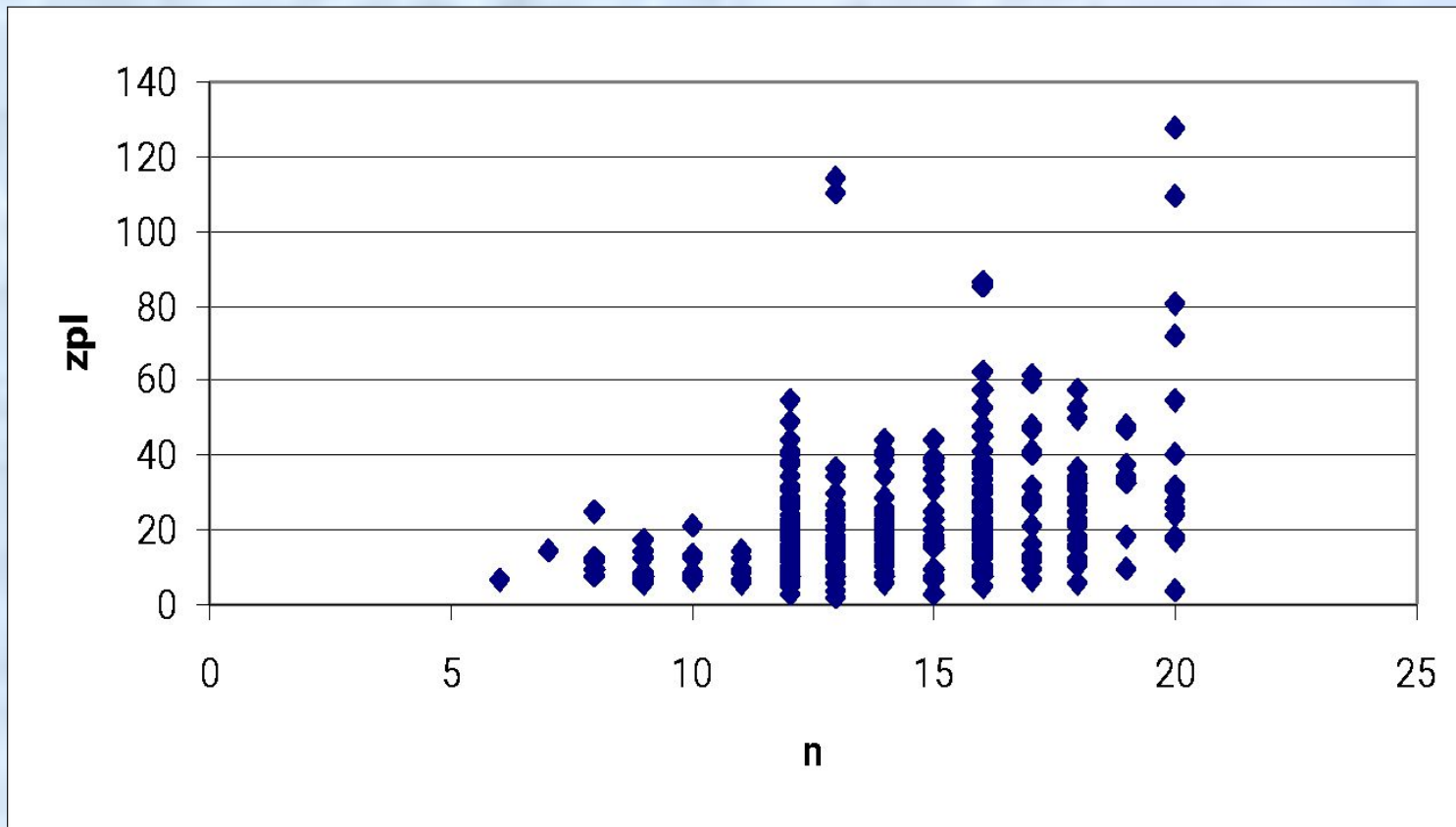
$$\hat{Y} = 0,9361X + 3,423$$

$$\bar{A} = 1,08\%$$

$$e = Y - \hat{Y}$$

$$a = \text{abs}(e)/y$$

ПРИМЕР УРАВНЕНИЯ ПАРНОЙ РЕГРЕССИИ



Данные 2002 г. о часовой заработной плате (\$ США) и уровне образования (лет) по 540 респондентам из национального опроса в США.

12 лет – средняя школа

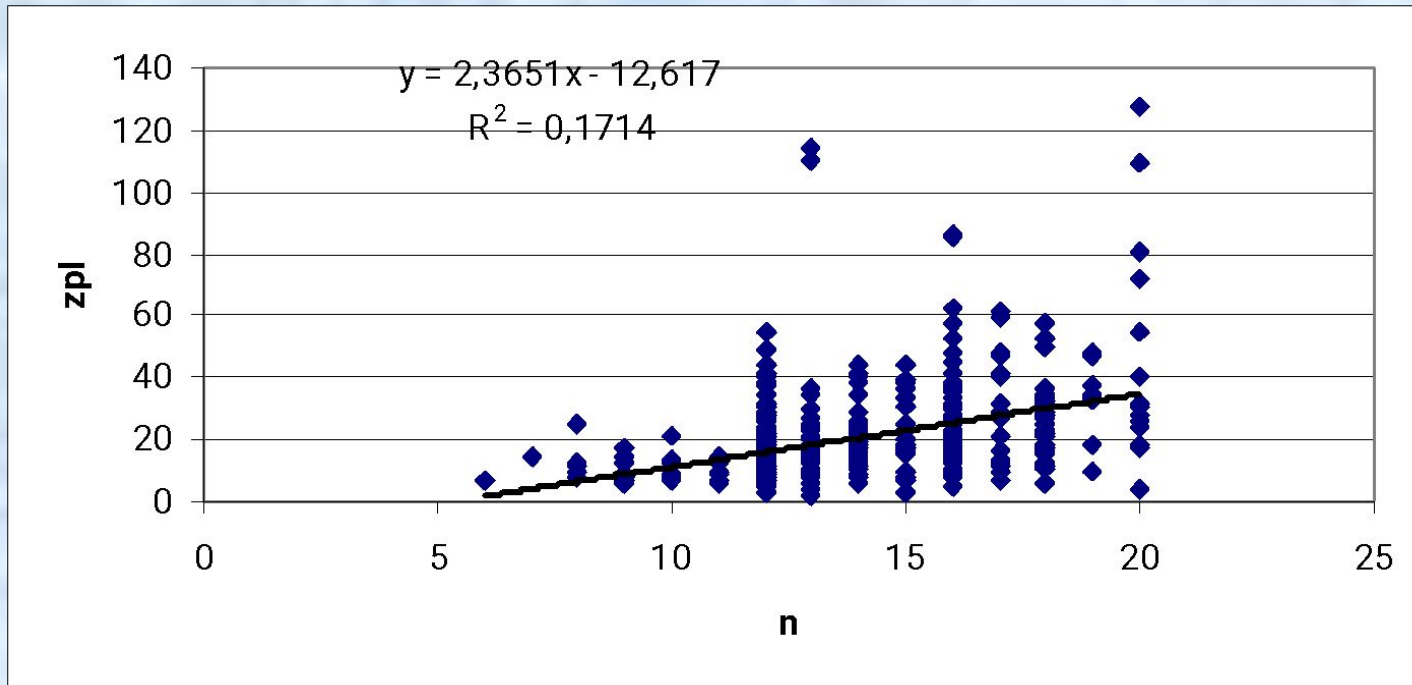
13-16 лет – колледж (бакалавриат)

17-18 лет – университет (магистратура)

19-20 лет - PhD

ПРИМЕР УРАВНЕНИЯ ПАРНОЙ РЕГРЕССИИ

$$Z_{pl} = -12,6188 + 2,3651N$$



Увеличению уровня образования на один год приведет в среднем к увеличению почасовой заработной платы на \$2.37

ПРИМЕР УРАВНЕНИЯ ПАРНОЙ РЕГРЕССИИ

ВЫВОД ИТОГОВ					
Регрессионная статистика					
Множественный R	0,413988				
R-квадрат	0,171386				
Нормированный R	0,169846				
Стандартная ошибка	13,30384				
Наблюдения	540				
Дисперсионный анализ					
	df	SS	MS	F	значимость
Регрессия	1	19695,21	19695,21	111,2774	8,96E-24
Остаток	538	95221,75	176,9921		
Итого	539	114917			
Коэффициенты регрессии					
Y-пересечение	-12,6168	3,118382	-4,04596	5,97E-05	-18,7425
n	2,365094	0,224205	10,54881	8,96E-24	1,92467

$$R^2 = 0,17$$

$$s = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - ax_i - b)^2}$$

$$Z_{pl} = -12,6168 + 2,3651N$$

zpl	прогноз	e	A
9	15,76428	-6,764278445	75,16%
10	15,76428	-5,764278445	57,64%
20,5	15,76428	4,735721555	23,10%
9,97	15,76428	-5,794278445	58,12%
12,3	15,76428	-3,464278445	28,16%
12,5	15,76428	-3,264278445	26,11%
15	15,76428	-0,764278445	5,10%
13,71	25,22465	-11,51465337	83,99%
27,32	15,76428	11,55572156	42,30%
11	18,12937	-7,129372177	64,81%
14,42	18,12937	-3,709372177	25,72%
14,32	15,76428	-1,444278445	10,09%
18,18	15,76428	2,415721555	13,29%
25,3	25,22465	0,075346628	0,30%
		Среднее	56,51%

$$Zpl = -12,6188 + 2,3651N$$

$e = zpl - \text{прогноз}$

$$a = \frac{\text{abs}(e)}{zpl} * 100$$

$$\bar{A} = 56,51\%$$