رگرسیون خطی ساده

An example

سن و میزان کلسترل 18 فرد
پرسیده شده است.

ID	Age	Chol (mg/ml)
1	46	3.5
2	20	1.9
3	52	4.0
4	30	2.6
5	57	4.5
6	25	3.0
7	28	2.9
8	36	3.8
9	22	2.1
10	43	3.8
11	57	4.1
12	33	3.0
13	22	2.5
14	63	4.6
15	40	3.2
16	48	4.2
17	28	2.3
18	49	4.0

سوال مورد نظر

- آیا ارتباطی بین سن و سطح کلسترل وجود دارد؟
- این رابطه چقدر قوی است؟
- برای یک سن مشخص میزان کلسترل چقدر پیش بینی میشود؟

همبستگی و رگرسیون

واریانس و کواریانس

$$\mathrm{var}(x) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} \qquad \mathrm{var}(y) = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{n-1}$$

- اندازه ارتباط بین دو متغیر اینگونه محاسبه میشود:

- Algebraically:

  $\mathrm{var}(x \pm y) = \mathrm{var}(x) + \mathrm{var}(y)$

  $\mathrm{var}(x \pm y) = \mathrm{var}(x) + \mathrm{var}(y) \pm 2\mathrm{cov}(x,y)$

Where: $\mathrm{cov}(x,y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

مفهوم واریانس و کواریانس

- واریانس همیشه مثبت است.
- اگر کواریانس دو متغیر صفر شد، دو متغیر مستقل هستند.
- کواریانس میتواند مثبت و یا منفی باشد.
- کواریانس منفی یعنی انحرافات دو متغیر در دو سمت مختلف حرکت میکنند.
- کواریانس مثبت یعنی انحرافات دو متغیر در سمت موافق حرکت میکنند.
- کواریانس دو متغیر قوت رابطه را نشان میدهد.

کواریانس و همبستگی

- کواریانس یک شاخص وابسته به واحد است.
- همبستگی بین دو متغیر از رابطه زیر بدست می آید.:

$$r = \frac{\mathrm{cov}(x,y)}{\sqrt{\mathrm{var}(x)\times \mathrm{var}(y)}} = \frac{\mathrm{cov}(x,y)}{SD_x \times SD_y}$$

ضریب همبستگی

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

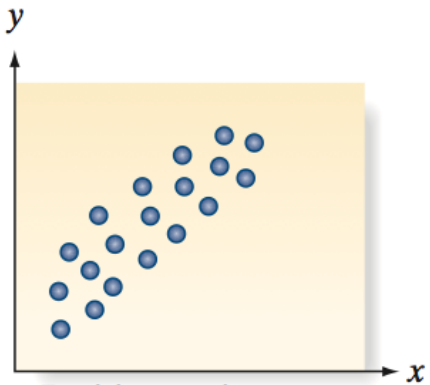
where

$$SS_{xy} = \sum \left(x - \bar{x}\right)\left(y - \bar{y}\right)$$

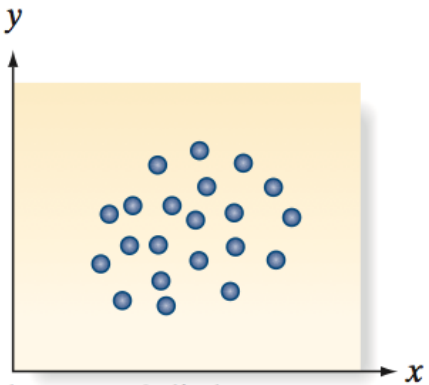
$$SS_{xx} = \sum \left(x - \bar{x}\right)^2$$

$$SS_{yy} = \sum \left(y - \bar{y}\right)^2$$

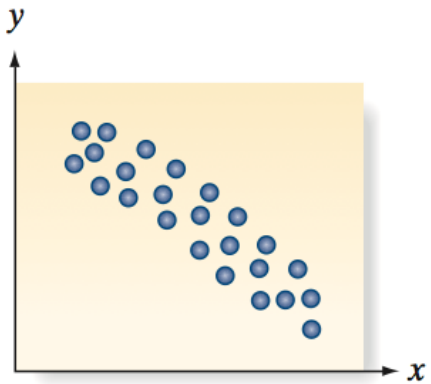
ضریب همبستگی



a. Positive r: y increases as x increases

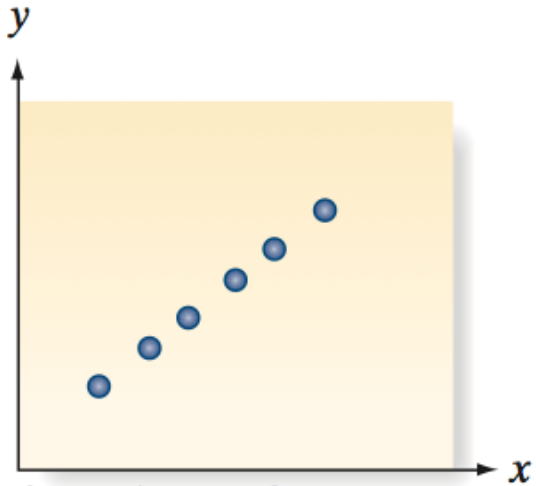


b. r near 0: little or no relationship between y and x

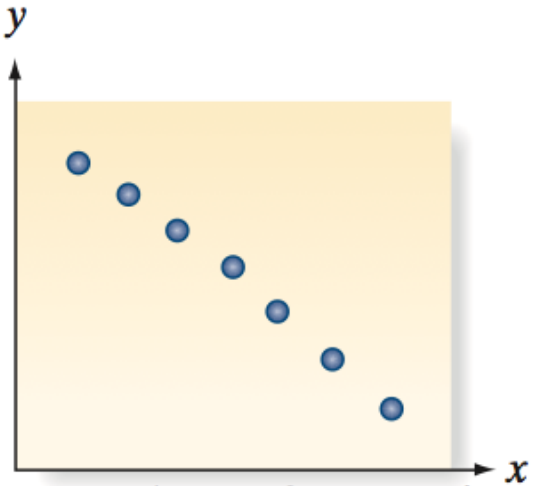


c. Negative r: y decreases as x increases

ضریب همبستگی

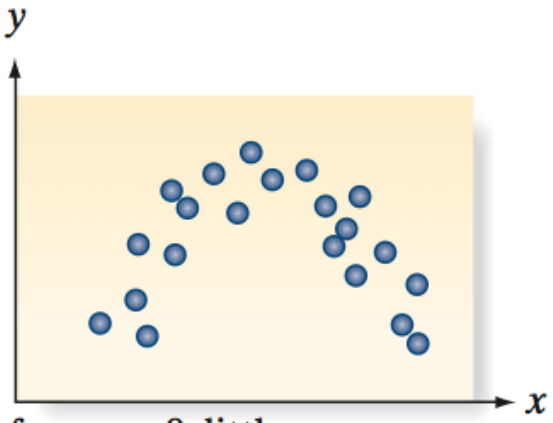


d. $r = 1$: a perfect positive relationship between y and x



e. $r = -1$: a perfect negative relationship between y and x

ضریب همبستگی



f. *r* near 0: little or no linear relationship between *y* and *x*

آزمون فرض همبستگی

- $H_o: r = 0$ versus $H_a: r \neq 0$.

- انحراف استاندارد برای ضریب همبستگی: $$SE(r) = \sqrt{\frac{1-r^2}{n-2}}$$

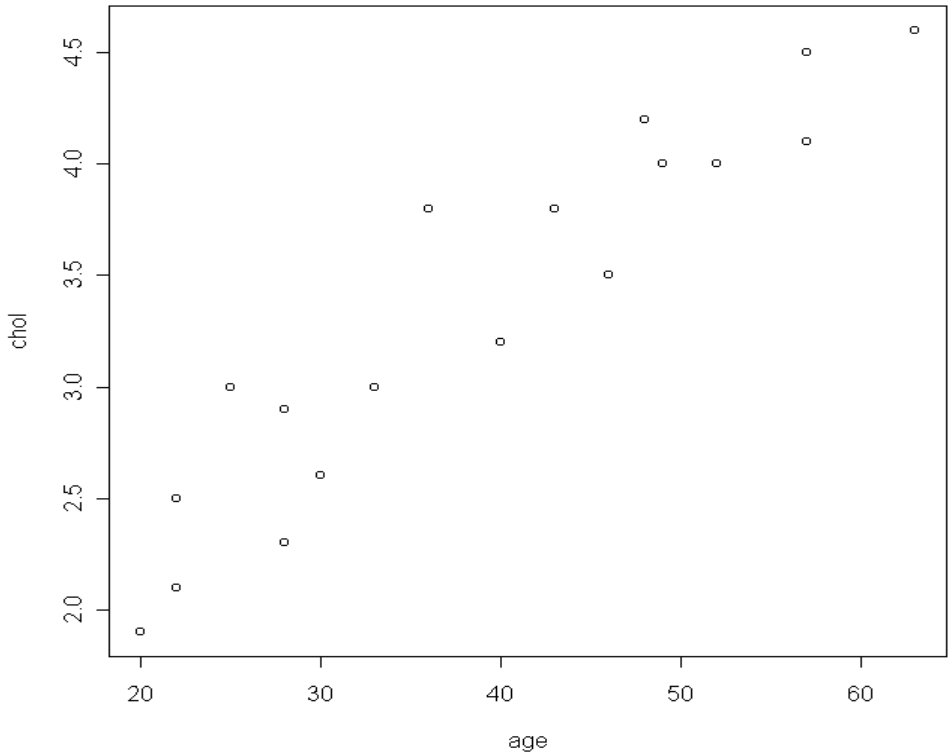
- آماره آزمون: $$t = r\sqrt{\frac{n-2}{1-r^2}}$$

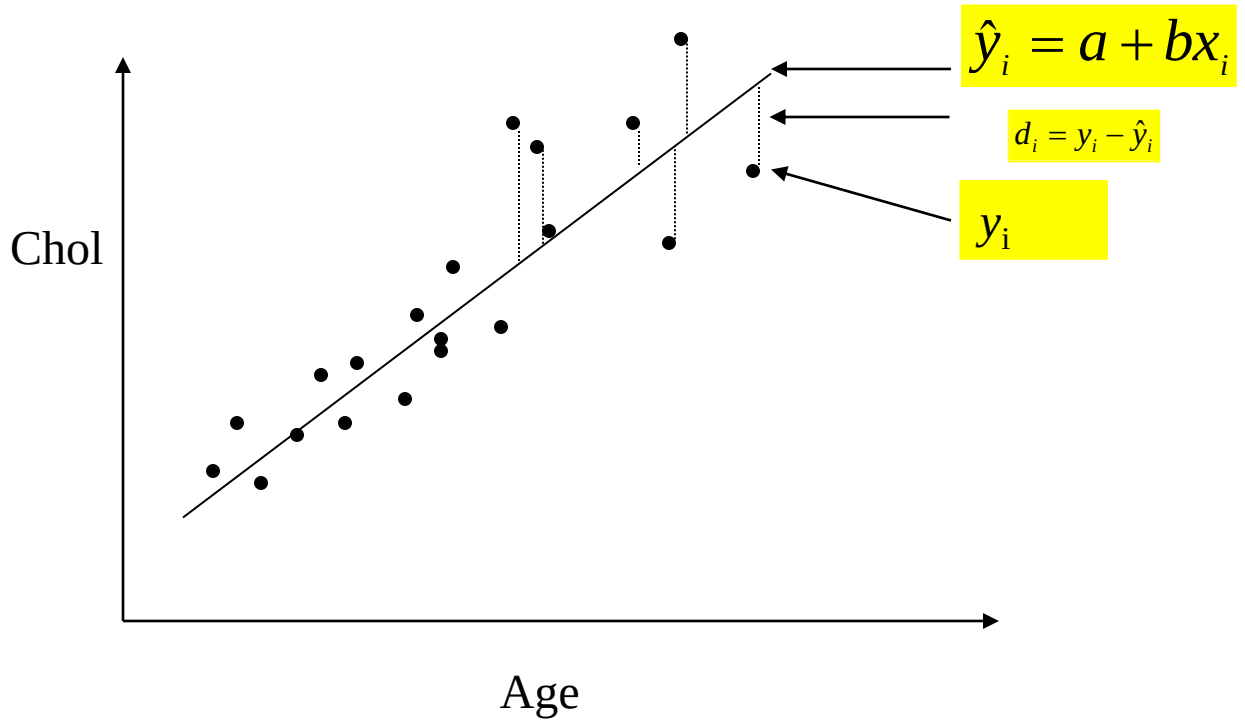
- این آماره دارای توزیع t با درجه آزادی n-2 میباشد.

رگرسیون خطی ساده

به طور کلی رگرسیون خطی ساده بهترین خط برای توصیف ارتباط بین دو متغیر را نشان میدهد. رگرسیون ساده یعنی تنها یک متغیر مستقل در مدل رگرسیونی ما وجود دارد.

رابطه بین سن و سطح کلسترل





یکی از روش های برازش خط رگرسیونی حداقل مربعات است.

رگرسیون خطی

- Y : متغیر پاسخ
- X : متغیر پیش بینی
- متغیر پاسخ حتما باید کمی ولی متغیر مستقل میتواند کمی و یا کیفی باشد.

  - مدل رگرسیونی

$$Y = \alpha + \beta X + \varepsilon$$

α : عرض از مبدا

B : شیب

E : خطای تصادفی که دارای توزیع نرمال با میانگین صفر و واریانس ثابت است.

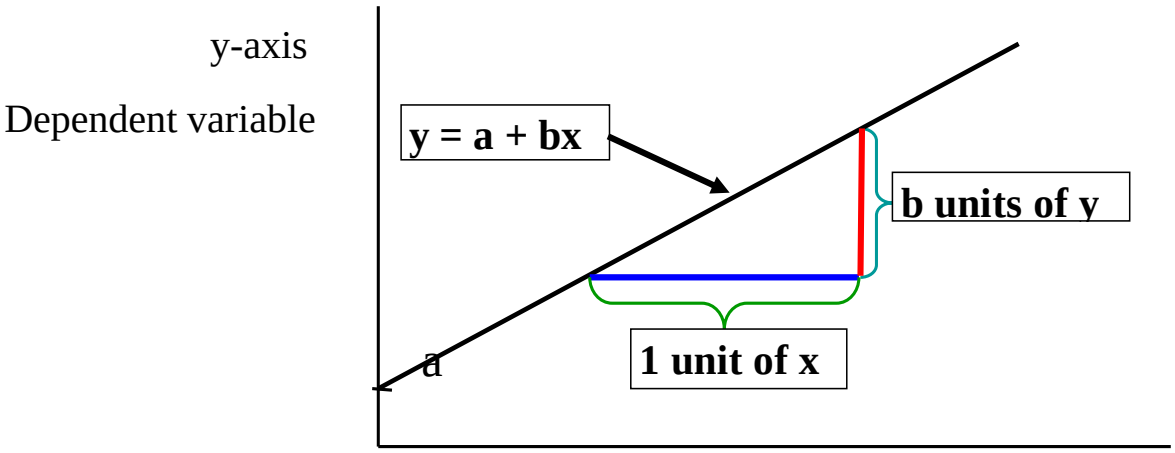
برآورد ضرایب رگرسیونی

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{S_{xy}}{S_{xx}}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$



a= مقدار متغیر پاسخ به ازای صفر بودن متغیر مستقل

b = مقدار تغییرات متغیر پاسخ به ازای یک واحد افزایش متغیر مستقل

آنالیز واریانس برای بررسی اثر گذار بودن متغیر مستقل در خط رگرسیونی

این آنالیز میسنجد که آیا مقدار شیب در خط رگرسیونی برابر صفر است یا خیر

1. The Hypothesis: $H_0: \beta = 0$ vs $H_1: \beta \neq 0$

2. The α level: $\alpha = 0.05$

3. The assumptions: Random normal samples for y-variable from populations defined by x-variable

4. The test statistic:

ANOVA				
Source	df	SS	MS	F
Regression	1	SS(*Reg*)	SS(*Reg*)/1	MS(*Reg*)/MS(*Res*)
Residual	n-2	SS(*Res*)	SS(*Res*)/(n-2)	
Total	n-1	SS(y)		

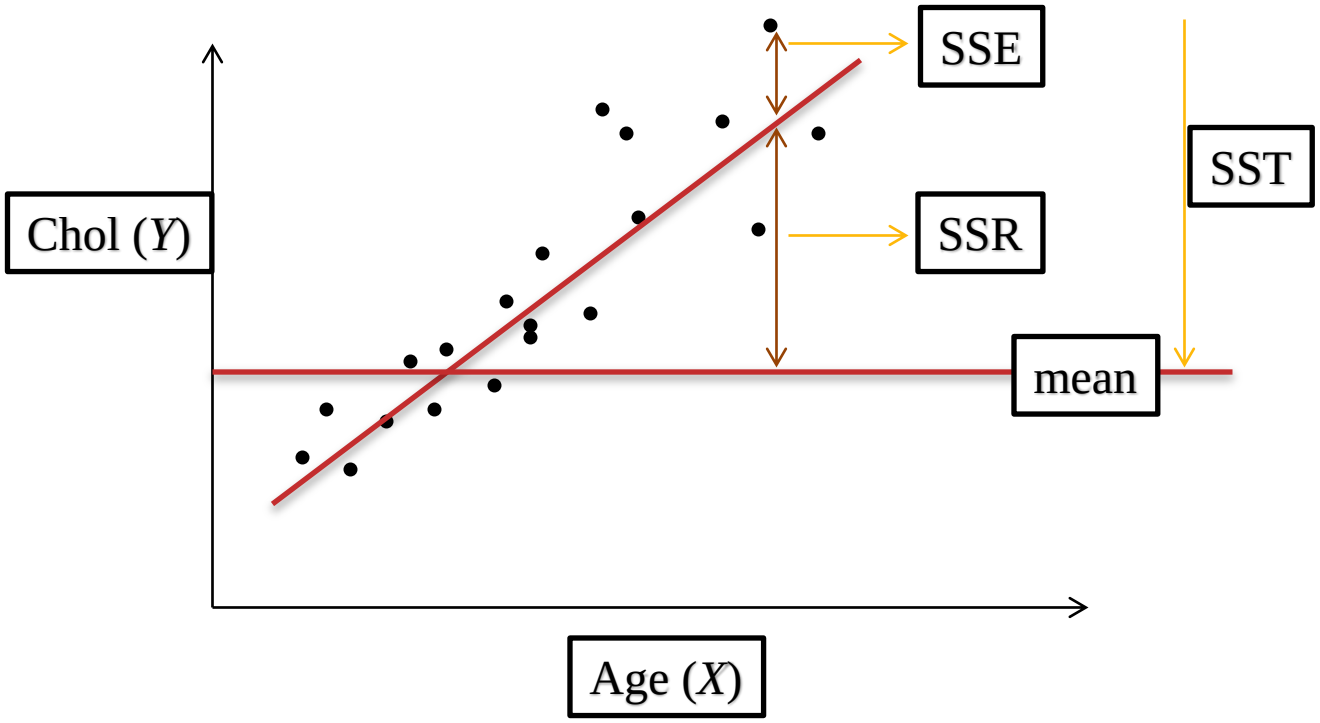
5. The rejection region : Reject $H_0: \beta = 0$ if the value calculated for F is greater than $F_{0.95}(1, n-2)$

نکته:

$R^2 = SSR / SST$

$$r = \hat{\beta}_1 \left(\frac{S_{xx}}{SS_T} \right)^{1/2}$$

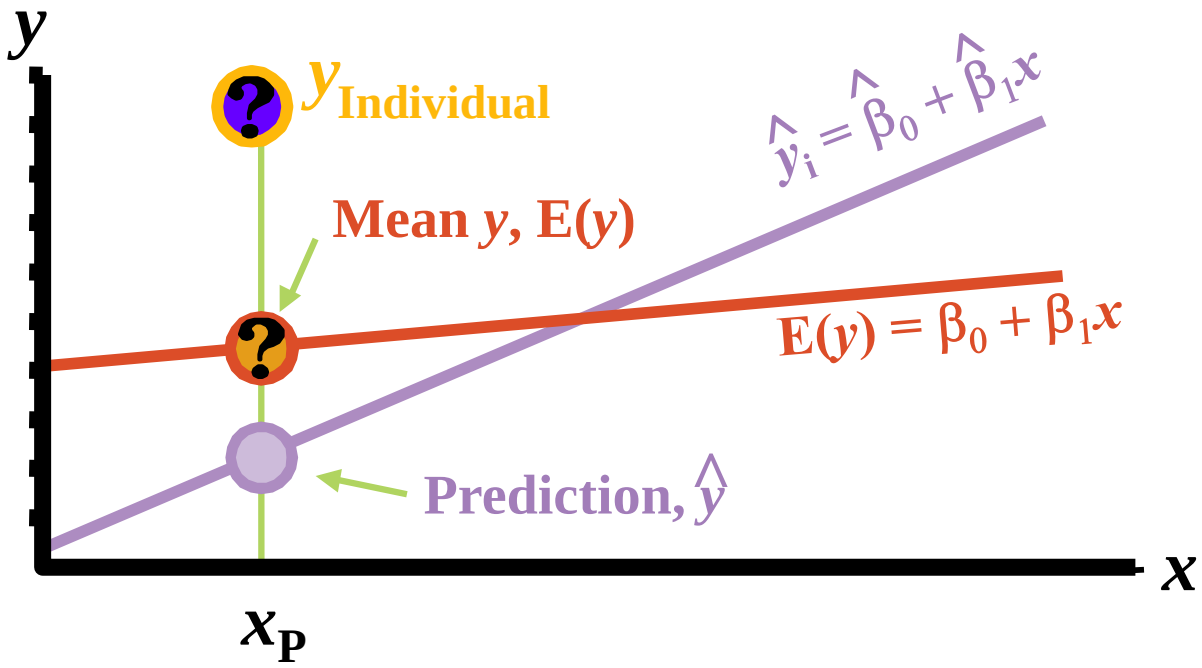
$$r_{xy} = (\text{sign of } b_1) \sqrt{r^2}$$


SSE
SST
SSR
Chol (Y)
mean
Age (X)

تجزیه تغییرات کل

- Some statistics:
  - Total variation: $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$
  - Attributed to the model: $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$
  - Residual sum of square: $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
  - $SST = SSR + SSE$
  - $SSR = SST - SSE$

مقدار پیش بینی از خط رگرسیون



مثال:

داده های زیر مربوط به فشارخون دیاستولیک در زمان های مختلف بعد از درمان برای 5 فرد میباشد.

آیا رابطه ایی بین درمان و فشارخون وجود دارد؟

	Time		DPB		
Patient	x	x^2	y	y^2	xy
1	0	0	72	5,184	0
2	5	25	66	4,356	330
3	10	100	70	4,900	700
4	15	225	64	4,096	960
5	20	400	66	4,356	1,320
Sum	50	750	338	22,892	3,310
Mean	10		67.6		
n	5		5		

For the blood pressure data,

$$\bar{x} = 50/5 = 10,$$

$$\bar{y} = 338/5 = 67.6,$$

the slope is

$$b = \frac{\sum xy - \sum x \sum y / n}{\sum x^2 - (\sum x)^2 / n} = \frac{SS(xy)}{SS(x)},$$

$$b = \frac{3,310 - (50)(338)/5}{750 - (50)^2/5} = -0.28$$

and the intercept is

$$a = \bar{y} - b\bar{x},$$

$$a = 67.6 - (-0.28)10 = 70.4$$

The best line is

$$y = a + bx = 70.4 - 0.28x$$

ANOVA

Source	df	SS	MS	F
Regression	1	19.6	19.6	2.49
Residual	3	23.6	7.89	
Total	4	43.2		

$H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$

For $\alpha = 0.05$ $F_{0.95(1,3)} = 10.1$, Hence accept $H_0 : \beta = 0$

$$R^2 = \frac{SS(Regression)}{SS(Total)} = \frac{19.6}{43.2} = 0.4537 \quad \text{or} \quad 45.37\%$$

Note: The above hypothesis test does not asses how well the straight line fits the data.