least squares line

- "least squares"
- slope and intercept

Dr. Mine Çetinkaya-Rundel
Duke University

a measure for the best line

Option 1: Minimize the sum of magnitudes (absolute values) of residuals

$$|e_1| + |e_2| + \cdots + |e_n|$$

✔ **Option 2:** Minimize the sum of squared residuals – least squares

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

why least squares?

- most commonly used
- easier to compute by hand and using software
- in many applications, a residual twice as large as another is more than twice as bad

least squares line

$$\hat{y} = \beta_0 + \beta_1 x$$

predicted response

intercept

slope

explanatory

notation

	parameter	point estimate
intercept	β_0	b_0
slope	β_1	b_1

estimating the regression parameters: slope

slope: $b_1 = \frac{s_y}{s_x} R$

s_x : SD of x
s_y : SD of y
$R = cor(x, y)$

practice

The standard deviation of % living poverty is 3.1%, and the standard deviation of % HS graduates is 3.73%. Given that the correlation between these variable is -0.75, what is the slope of the regression line for predicting % living poverty from % HS graduates?

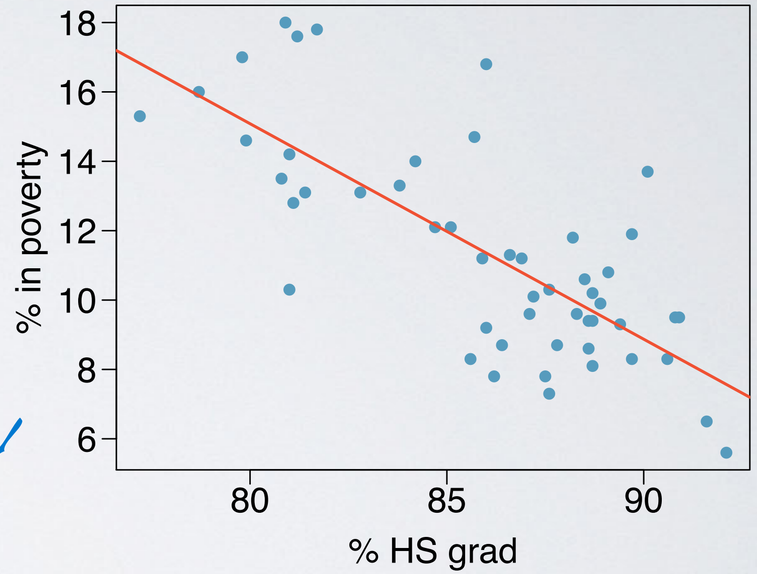
$s_y = 3.1\%$

$s_x = 3.73\%$

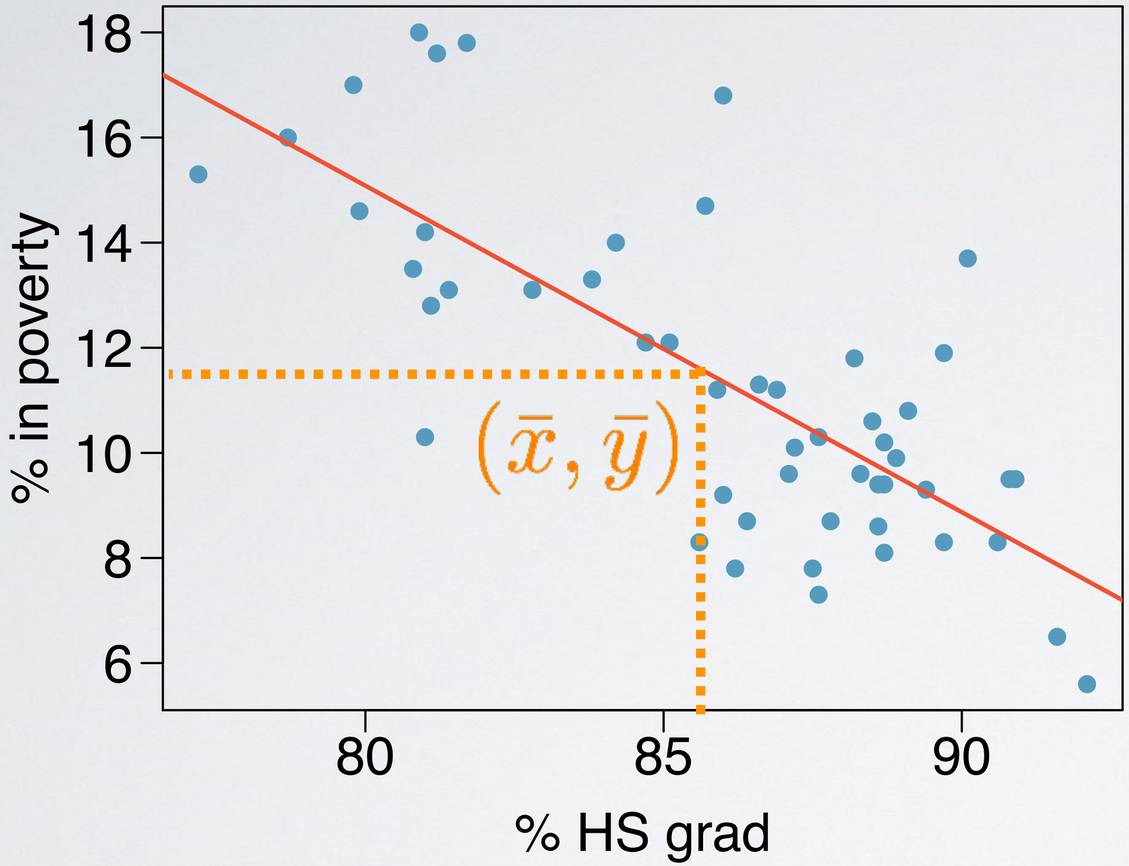
$R = -0.75$

$$b_1 = \frac{s_y}{s_x} R = \frac{3.1}{3.73} \times -0.75 \approx -0.62$$

For each % point increase in HS graduate rate, we would expect the % living in poverty to be lower on average by 0.62% points.



estimating the regression parameters: intercept



the least squares line always goes through $(\bar{x}, \bar{y})$

$$\bar{y} \quad \hat{y} = b_0 + b_1 x \quad \bar{x}$$

intercept: $b_0 = \bar{y} - b_1 \bar{x}$

practice

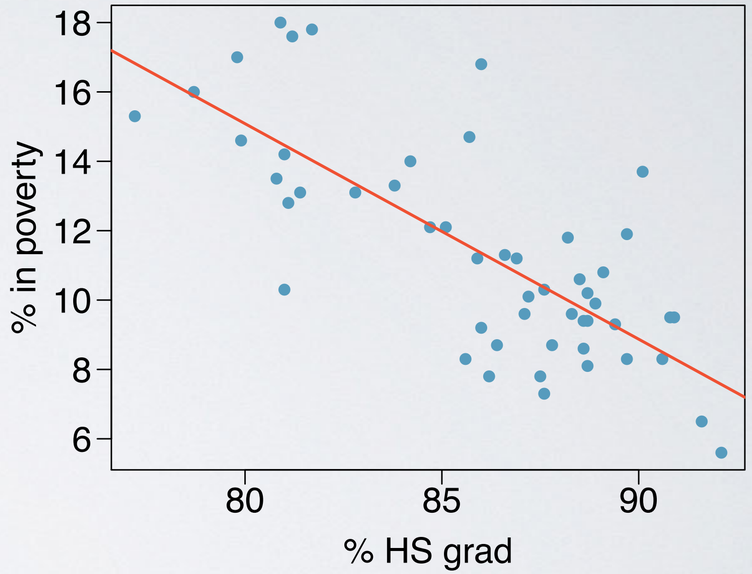
Given that the average % living in poverty is 11.35%, and the average % HS graduates is 86.01%, what is the intercept of the regression line for predicting % living poverty from % HS graduates?

$\bar{y} = 11.35\%$

$\bar{x} = 86.01\%$

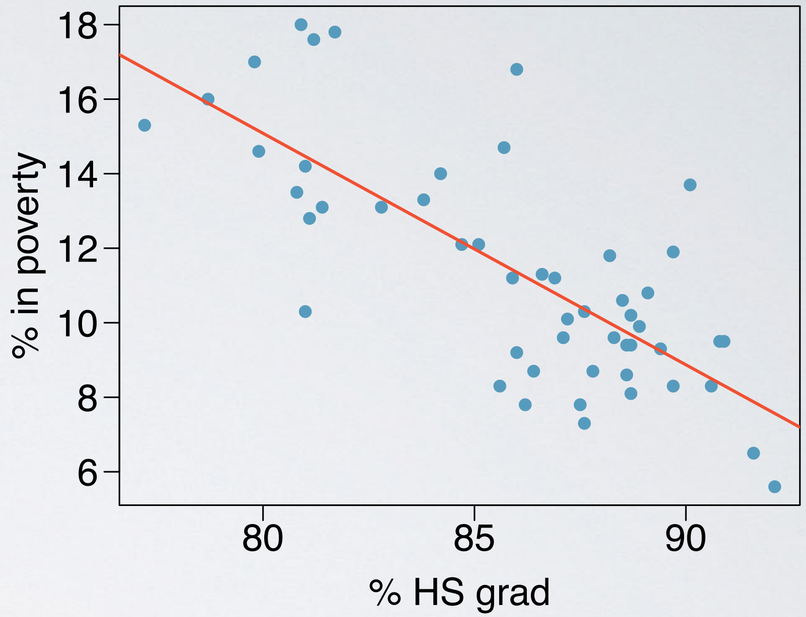
$b_0 = \bar{y} - b_1 \bar{x} = 11.35 - (-0.62)\ 86.01 = 64.68$

States with no HS graduates are expected on average to have 64.68% of their residents living below the poverty line.



$\widehat{\%\ \text{in poverty}} = 64.68 - 0.62 \times \%\ \text{HS grad}$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	64.78	6.80	9.52	0.00
hsgrad	-0.62	0.08	-7.86	0.00



recap

- intercept: When $x = 0$, y is expected to equal the intercept.
  - may be meaningless in context of the data, and only serve to adjust the height of the line
- slope: For each unit increase in x, y is expected to be higher/lower on average by the slope.