# Lecture 19 - Correlation and Regression

Sta102 / BME102

April 11, 2016

Colin Rundel

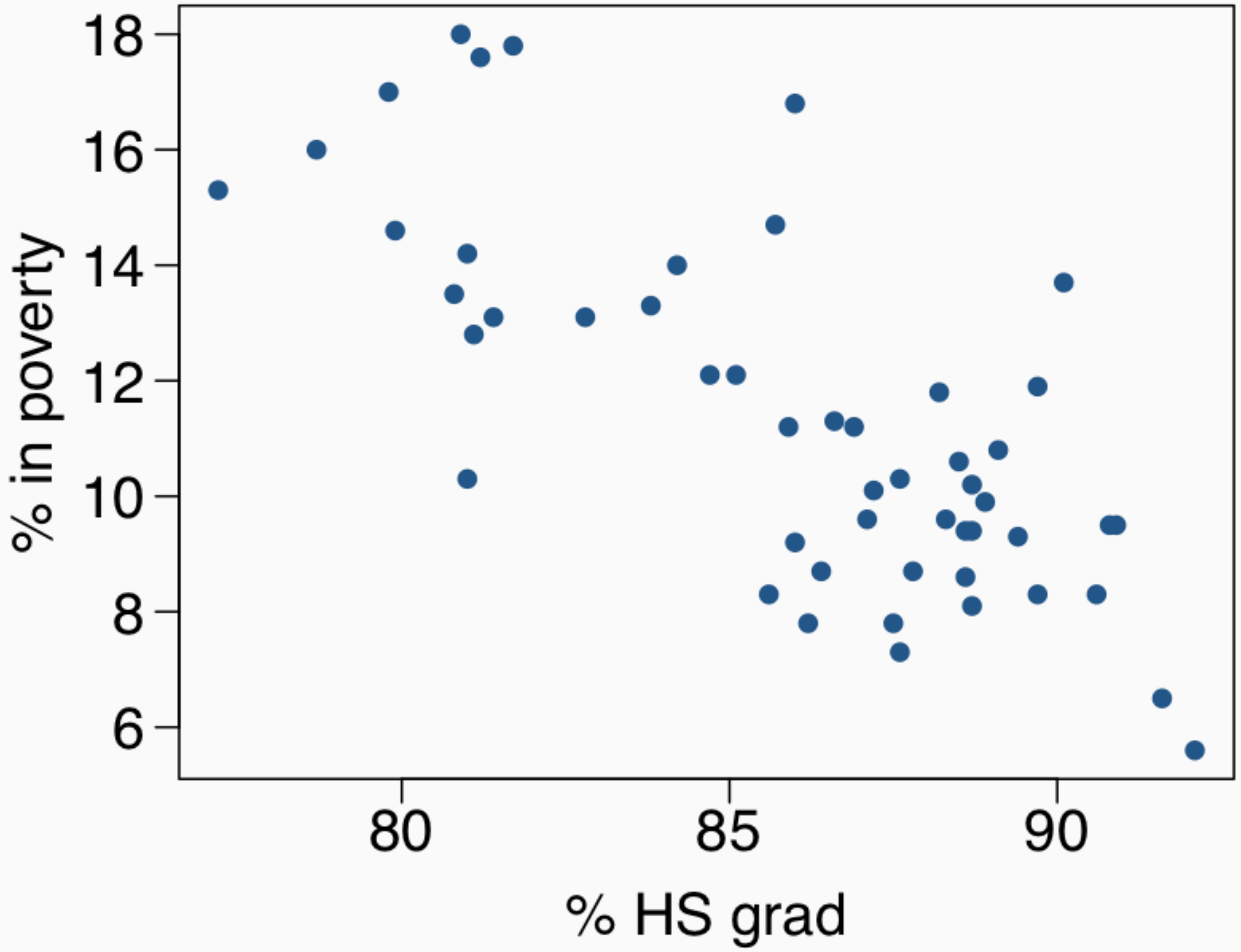
# Modeling numerical variables

# Modeling numerical variables

- So far we have worked with single numerical and categorical variables, and explored relationships between numerical and categorical, and two categorical variables.
- Today we will learn to quantify the relationship between two numerical variables.
- Next week we will learn to model numerical variables using many predictor (independent) variables (including both numerical and categorical) at once.

# Poverty vs. HS graduate rate

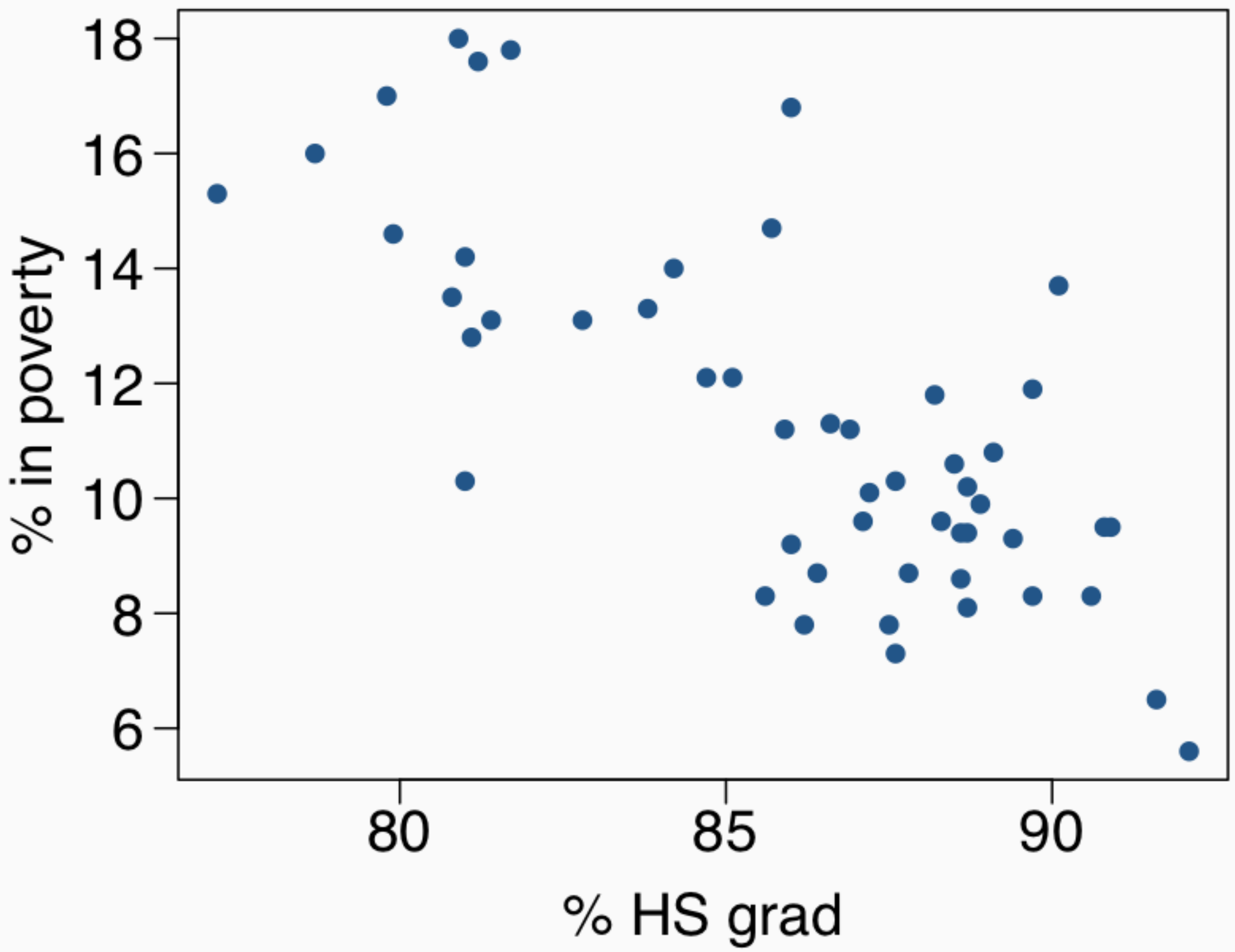
The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).




Response?

# Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).

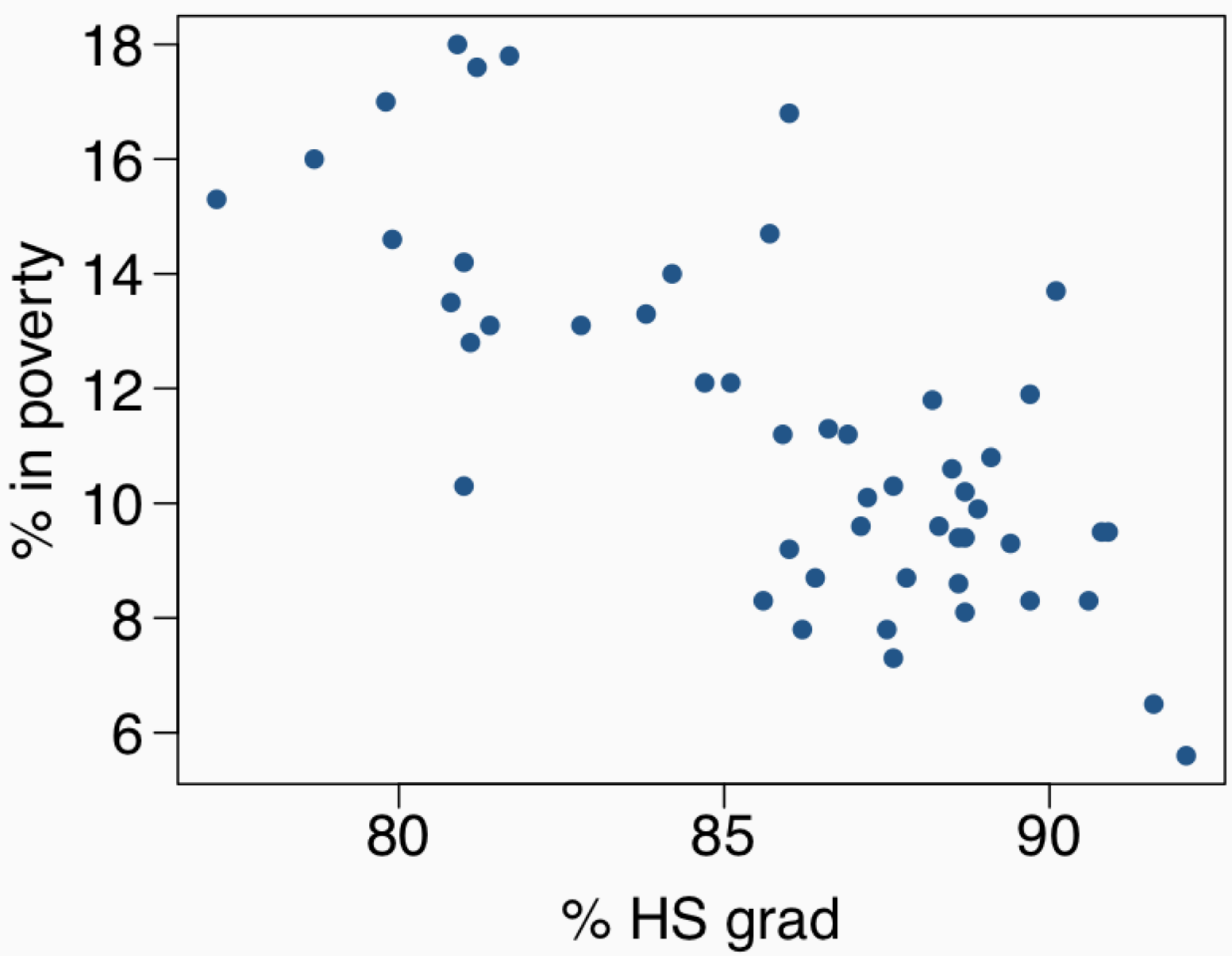



Response?

*% in poverty*

# Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).



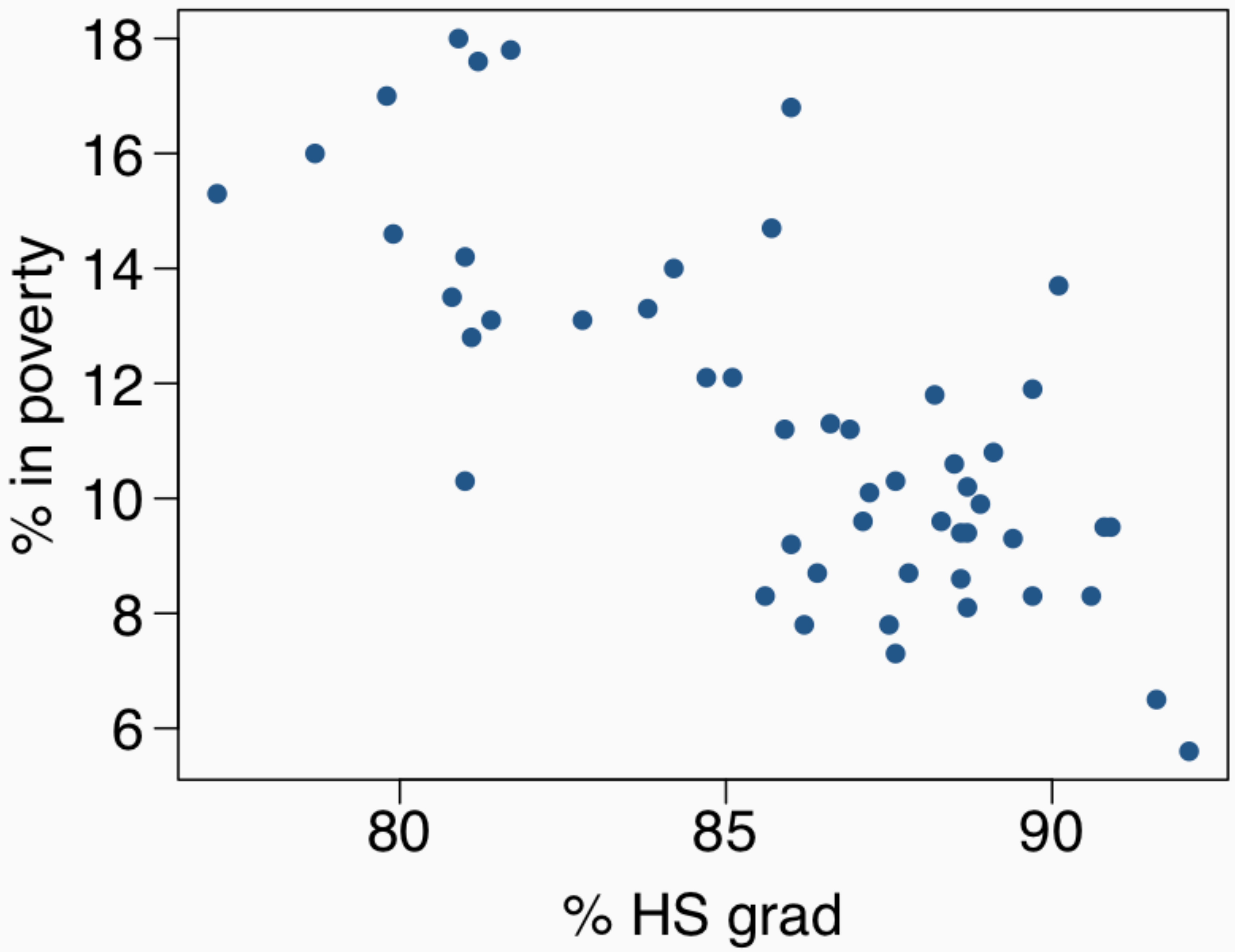

Response?

*% in poverty*

Predictor?

# Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).




Response? (Dep)

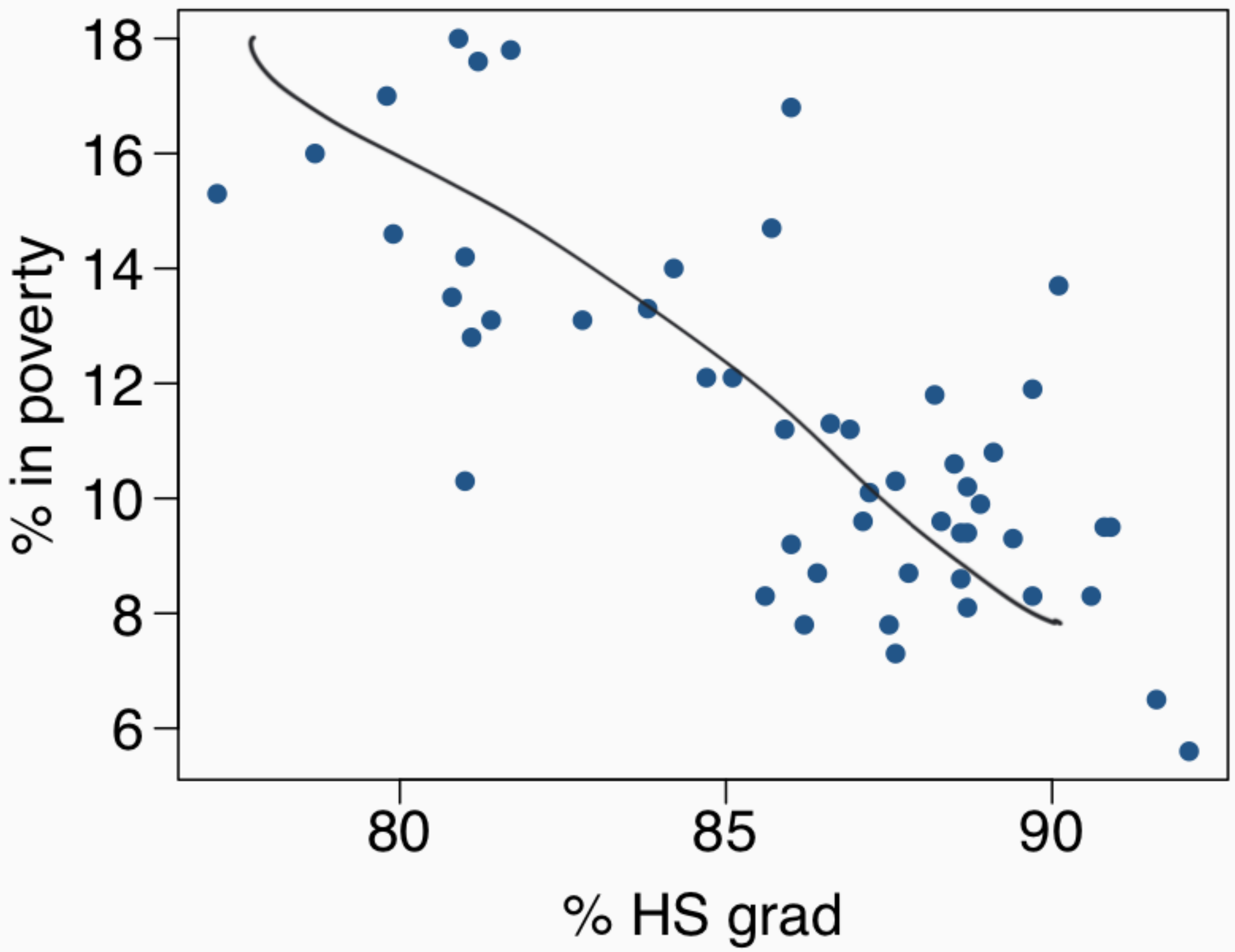
*% in poverty*

Predictor? (Ind)

*% HS grad*

# Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).




Response?

*% in poverty*

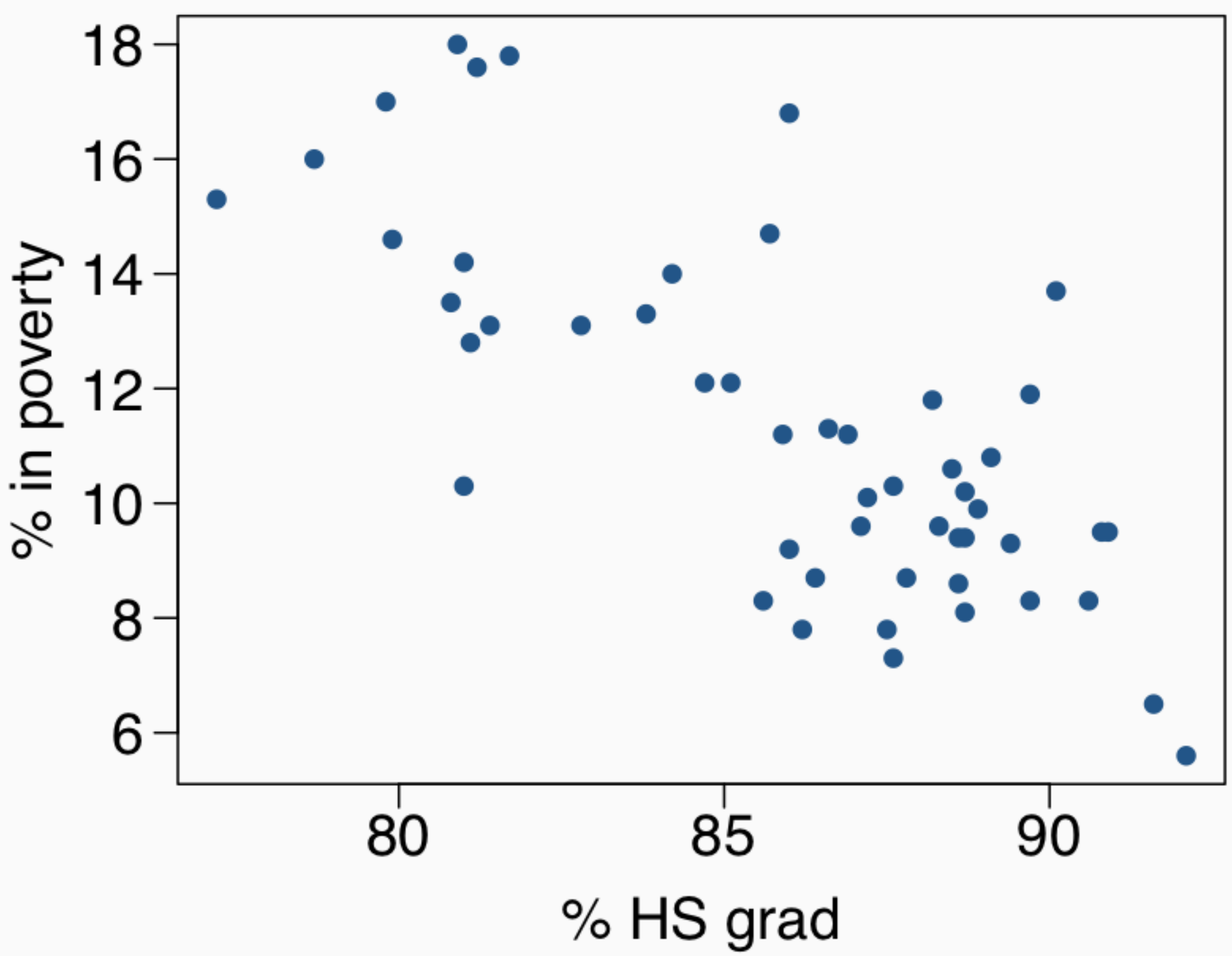
Predictor?

*% HS grad*

Relationship?

# Poverty vs. HS graduate rate

The *scatterplot* below shows the relationship between HS graduate rate in all 50 US states and DC and the % of residents who live below the poverty line (income below $23,050 for a family of 4 in 2012).




Response?

*% in poverty*

Predictor?

*% HS grad*

Relationship?

- *linear*
- *negative*
- *moderately strong*

# Covariance and Correlation

# Covariance

We have previously discussed variance as a measure of uncertainty of a sampled variable

$$Var(X) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)^2$$

we can generalize this to two variables,

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)$$

This quantity is called Covariance, and it is a measure of the degree to which $X$ and $Y$ tend to be large (or small) at the same time.

# Covariance, cont.

The magnitude of the covariance is not immediately useful as it is affected by the magnitude of both $X$ and $Y$.

However, the sign of the covariance tells us something useful about the relationship between $X$ and $Y$.

# Covariance, cont.

The magnitude of the covariance is not immediately useful as it is affected by the magnitude of both $X$ and $Y$.

However, the sign of the covariance tells us something useful about the relationship between $X$ and $Y$.

Consider the following conditions:

# Covariance, cont.

The magnitude of the covariance is not immediately useful as it is affected by the magnitude of both $X$ and $Y$.

However, the sign of the covariance tells us something useful about the relationship between $X$ and $Y$.

Consider the following conditions:

- $x_i > \mu_X$ and $y_i > \mu_Y$ then $(x_i - \mu_X)(y_i - \mu_Y)$ will be positive.

# Covariance, cont.

The magnitude of the covariance is not immediately useful as it is affected by the magnitude of both $X$ and $Y$.

However, the sign of the covariance tells us something useful about the relationship between $X$ and $Y$.

Consider the following conditions:

- $x_i > \mu_X$ and $y_i > \mu_Y$ then $(x_i - \mu_X)(y_i - \mu_Y)$ will be positive.
- $x_i < \mu_X$ and $y_i < \mu_Y$ then $(x_i - \mu_X)(y_i - \mu_Y)$ will be positive.

# Covariance, cont.

The magnitude of the covariance is not immediately useful as it is affected by the magnitude of both $X$ and $Y$.

However, the sign of the covariance tells us something useful about the relationship between $X$ and $Y$.

Consider the following conditions:

- $x_i > \mu_X$ and $y_i > \mu_Y$ then $(x_i - \mu_X)(y_i - \mu_Y)$ will be positive.
- $x_i < \mu_X$ and $y_i < \mu_Y$ then $(x_i - \mu_X)(y_i - \mu_Y)$ will be positive.
- $x_i > \mu_X$ and $y_i < \mu_Y$ then $(x_i - \mu_X)(y_i - \mu_Y)$ will be negative.

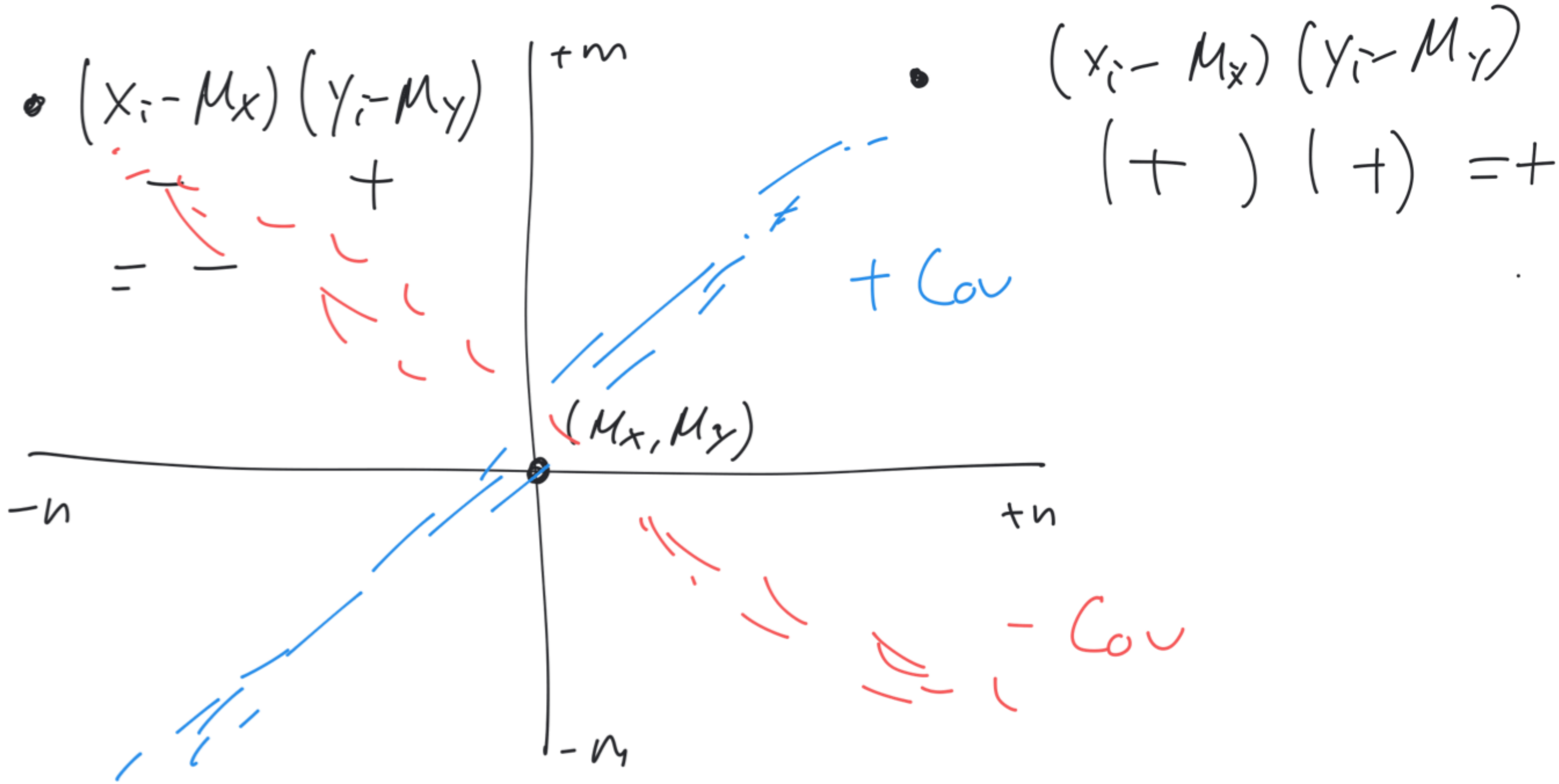
# Covariance, cont.

The magnitude of the covariance is not immediately useful as it is affected by the magnitude of both $X$ and $Y$.

However, the sign of the covariance tells us something useful about the relationship between $X$ and $Y$.

Consider the following conditions:

- $x_i > \mu_X$ and $y_i > \mu_Y$ then $(x_i - \mu_X)(y_i - \mu_Y)$ will be positive.
- $x_i < \mu_X$ and $y_i < \mu_Y$ then $(x_i - \mu_X)(y_i - \mu_Y)$ will be positive.
- $x_i > \mu_X$ and $y_i < \mu_Y$ then $(x_i - \mu_X)(y_i - \mu_Y)$ will be negative.
- $x_i < \mu_X$ and $y_i > \mu_Y$ then $(x_i - \mu_X)(y_i - \mu_Y)$ will be negative.



(xi − μx)(yi − μy)
− +
= −
+m
−n
+n
−m
(μx, μy)
(xi − μx)(yi − μy)
(+)(+) = +
+ Cov
− Cov

# Properties of Covariance

- $Cov(X, X) = Var(X)$
- $Cov(X, Y) = Cov(Y, X)$
- $Cov(X, Y) = 0$ if $X$ and $Y$ are independent
- $Cov(X, c) = 0$
- $Cov(aX, bY) = ab\ Cov(X, Y)$
- $Cov(X + a, Y + b) = Cov(X, Y)$
- $Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$

# Correlation

Since $Cov(X, Y)$ depends on the magnitude of $X$ and $Y$ we prefer to have a measure of association that is independent of the scale of the variables.

# Correlation

Since $Cov(X, Y)$ depends on the magnitude of $X$ and $Y$ we prefer to have a measure of association that is independent of the scale of the variables.

The most common measure of *linear* association is correlation, which is defined as

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \, \sigma_Y}$$

# Properties of Correlation

- *Correlation* describes the strength of the *linear* association between two variables.

# Properties of Correlation

- *Correlation* describes the strength of the *linear* association between two variables.
- It takes values between -1 (perfect negative) and +1 (perfect positive).

# Properties of Correlation

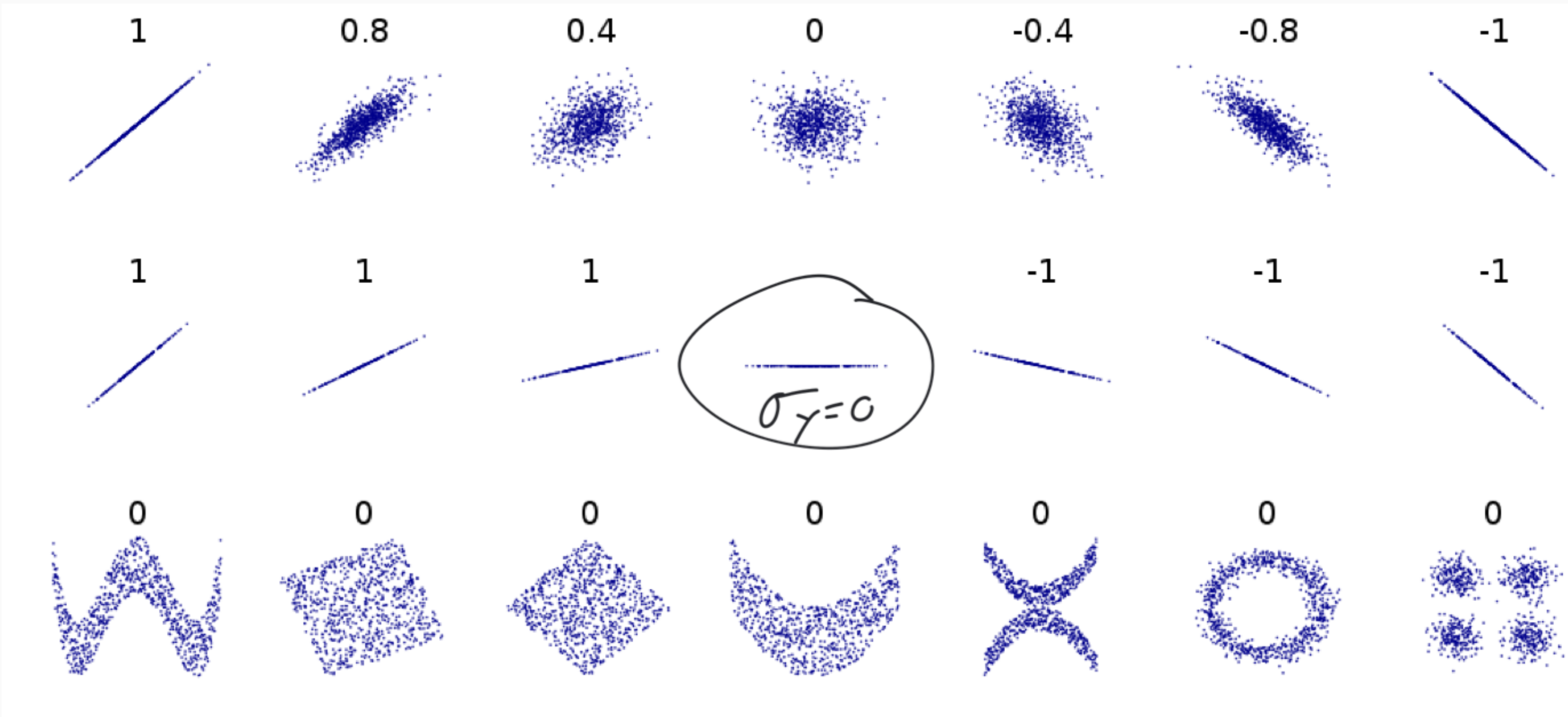
- *Correlation* describes the strength of the *linear* association between two variables.
- It takes values between -1 (perfect negative) and +1 (perfect positive).
- A value of 0 indicates no *linear* association.

# Properties of Correlation

- *Correlation* describes the strength of the *linear* association between two variables.
- It takes values between -1 (perfect negative) and +1 (perfect positive).
- A value of 0 indicates no *linear* association.
- We use $\rho$ to indicate the population correlation coefficient, and $R$ or $r$ to indicate the sample correlation coefficient.

$$R = \frac{Cov(x, y)}{s_x s_y}$$

# Correlation Examples




From http://en.wikipedia.org/wiki/Correlation

# Correlation and Independence

Given random variables $X$ and $Y$

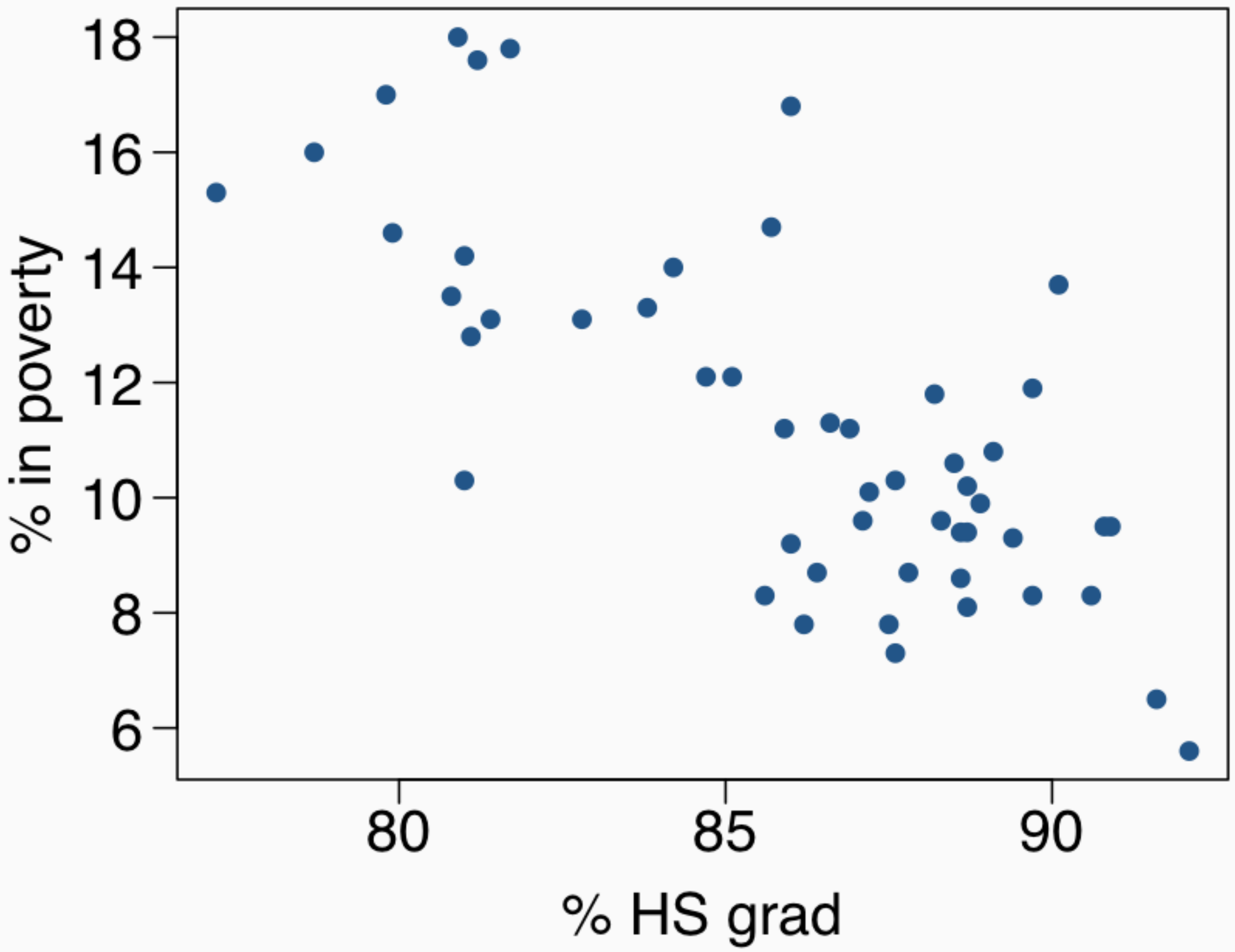
$$\text{If } X \text{ and } Y \text{ are independent} \implies Cov(X, Y) = \rho(X, Y) = 0$$

$$\text{If } Cov(X, Y) = \rho(X, Y) = 0 \not\Longrightarrow X \text{ and } Y \text{ are independent}$$

$\rho(X, Y) = 0$ *is necessary but not sufficient for independence.*

# Guessing the correlation

Which of the following is the best guess for the correlation between % in poverty and % HS grad?




(a) 0.6

(b) -0.75

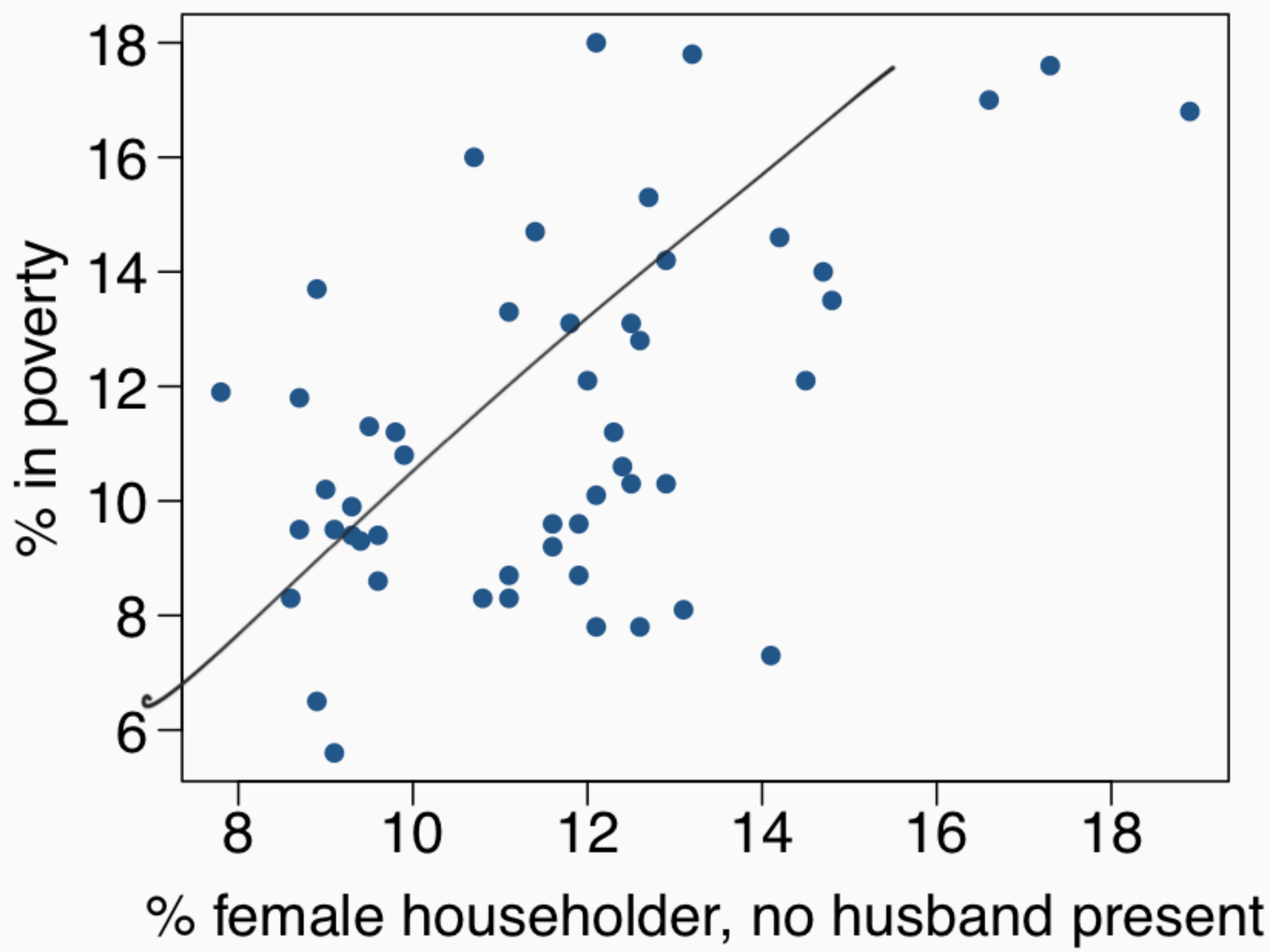
(c) -0.1

(d) 0.02

(e) -1.5

# Guessing the correlation

Which of the following is the best guess for the correlation between % in poverty and % single mother household?




(a) 0.1

(b) -0.6

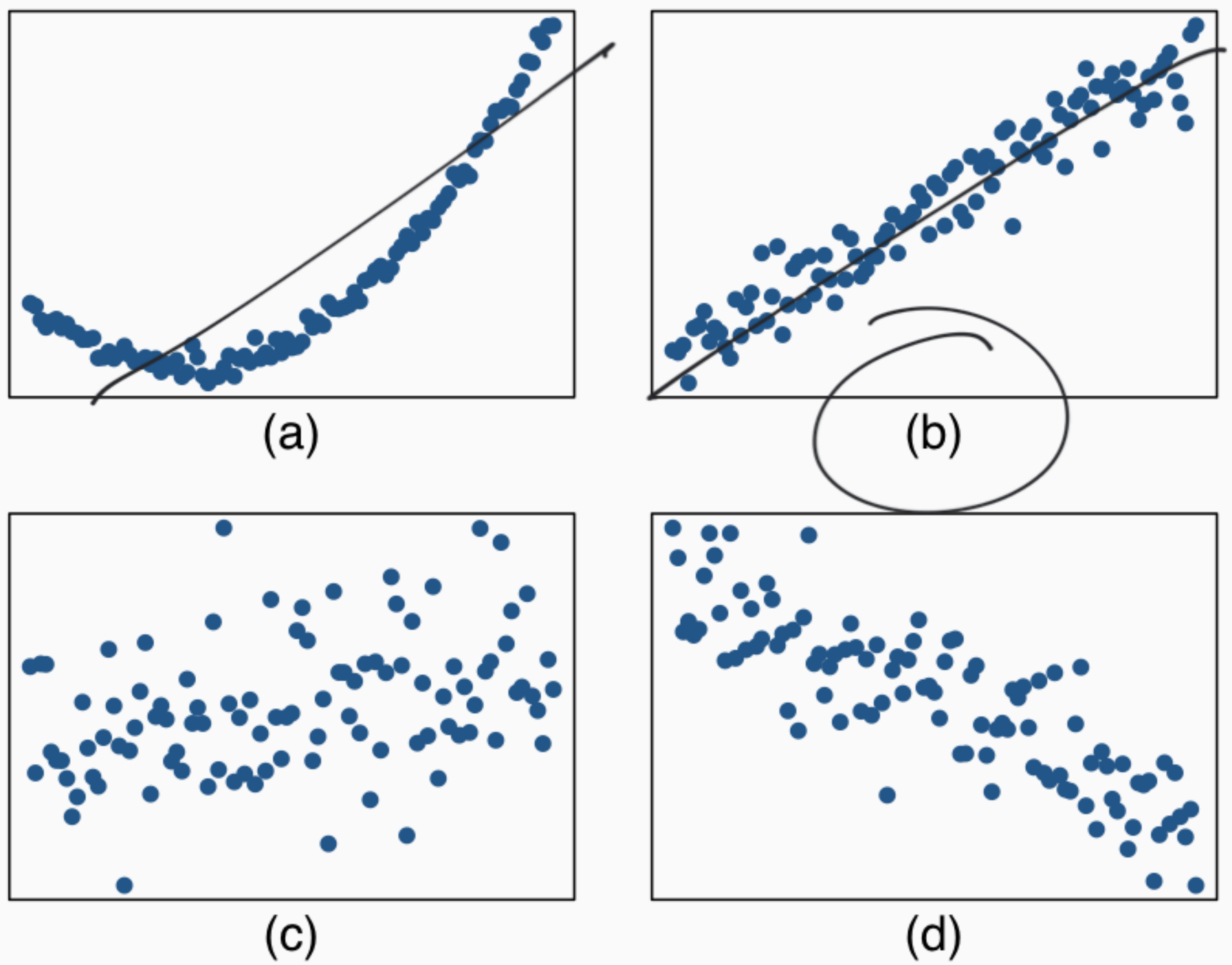
(c) -0.4

(d) 0.9

(e) 0.5

# Assessing the correlation

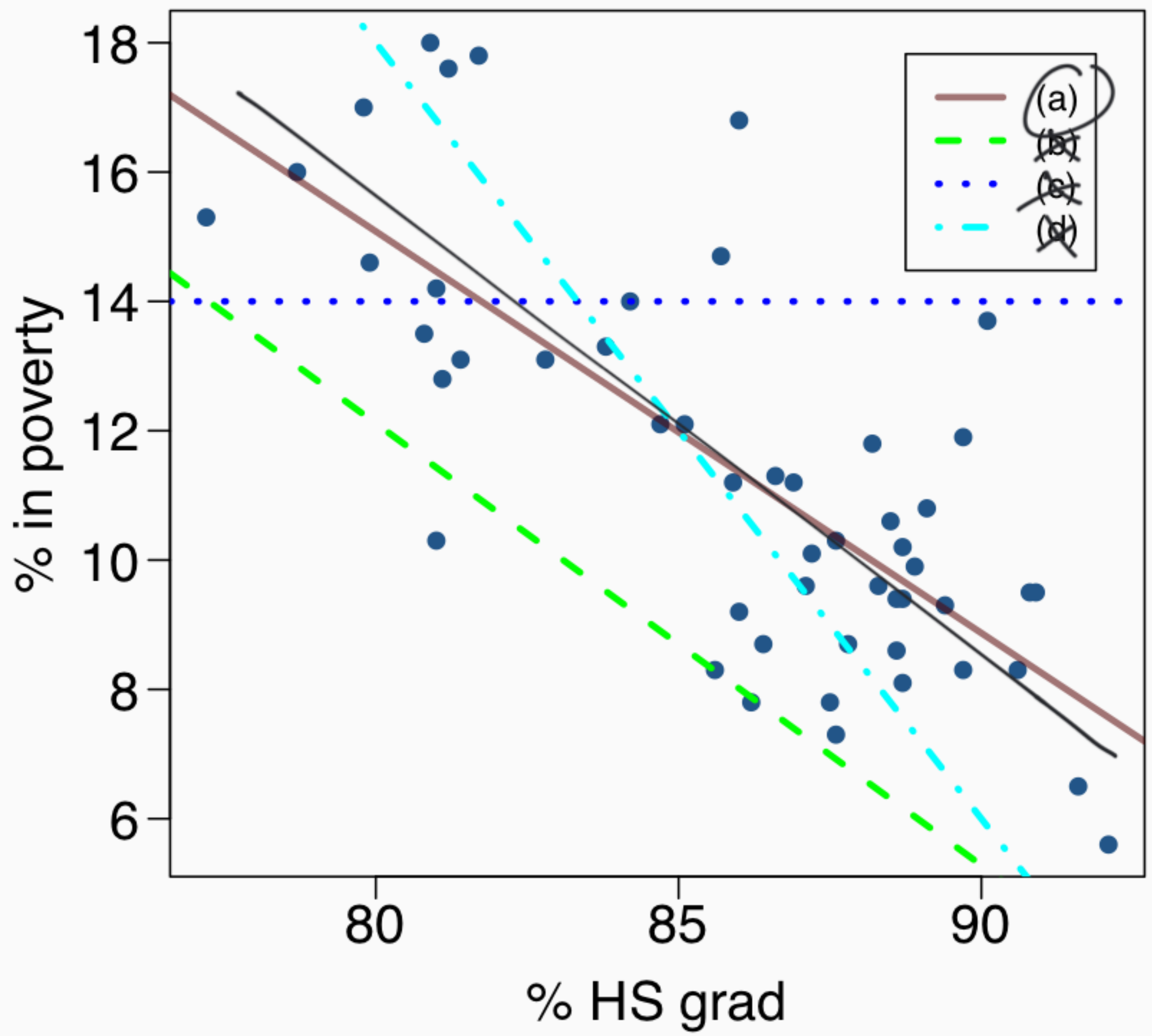
Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?



# Best fit line - least squares regression

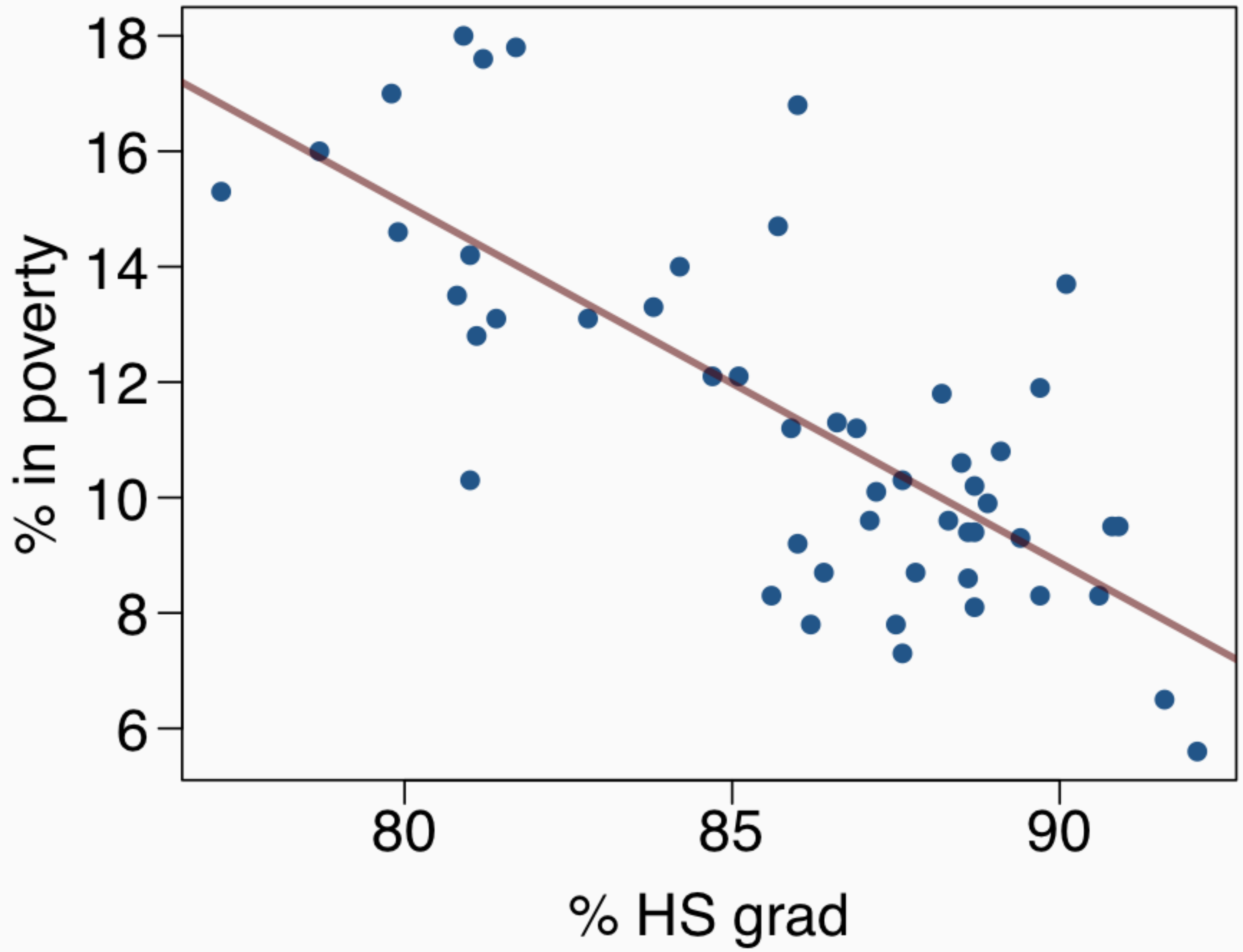
# Eyeballing the line

Which of the following appears to be the line that best fits the linear relationship between % in poverty and % HS grad?



# Line Equation

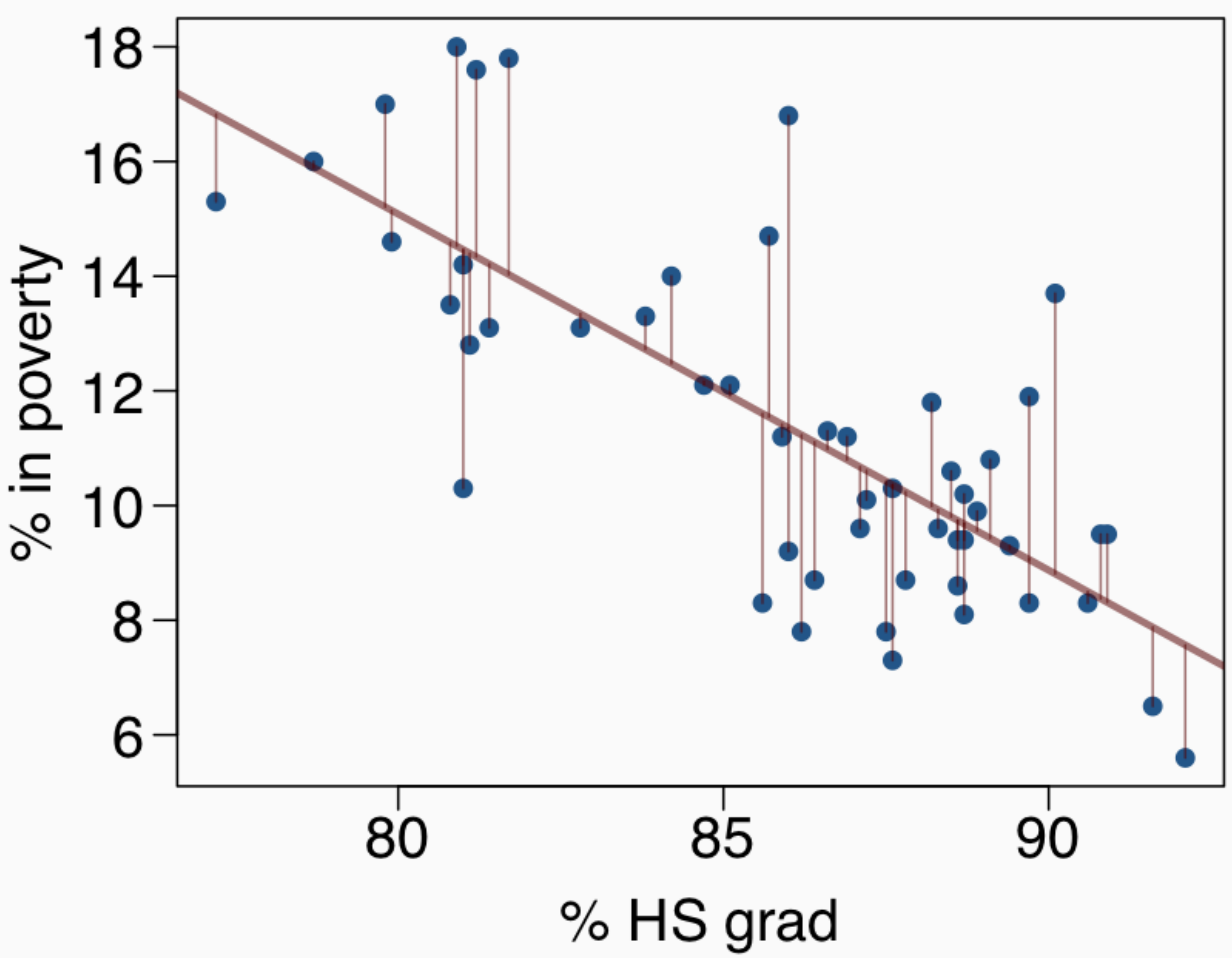
The line shown can be described by an equation of the form $\hat{y}_i = \beta_0 + \beta_1 x_i$, we would like a measure of the quality of its fit.



# Residuals

Just like with ANOVA, we can think about each value ($y_i$) as being the result of our model ($\hat{y}_i$) and some unexplained error ($e_i$) - this error is what we call a residual.

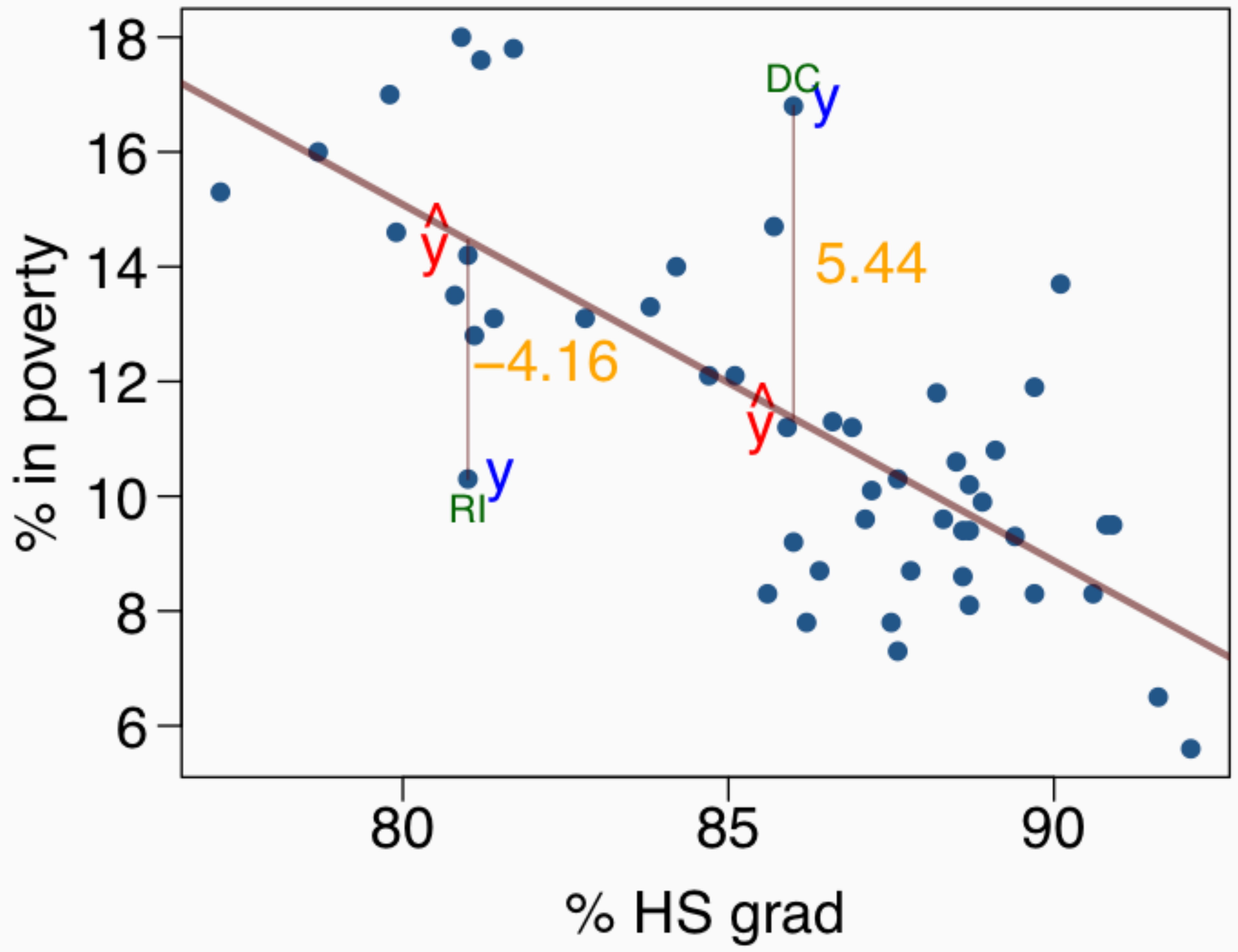
$$y_i = \hat{y}_i + e_i = \beta_0 + \beta_1 x_i + e_i$$



# Residual Examples

We can think about a residual being the difference between our observed outcome ($y_i$) minus our predicted outcome.

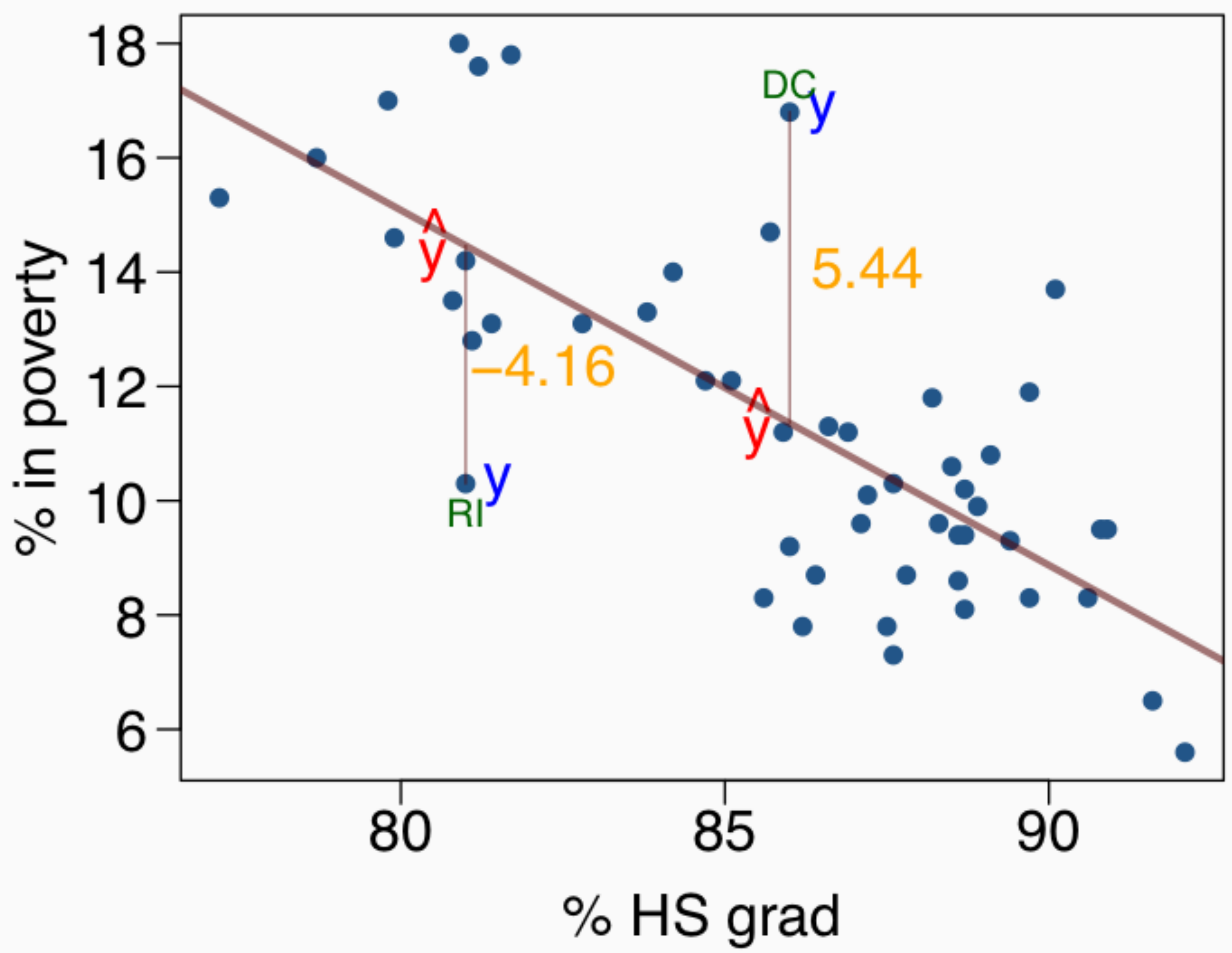
$$e_i = y_i - \hat{y}_i = y_i - \beta_0 - \beta_1 x_i$$



# Residual Examples

We can think about a residual being the difference between our observed outcome ($y_i$) minus our predicted outcome.

$$e_i = y_i - \hat{y}_i = y_i - \beta_0 - \beta_1 x_i$$

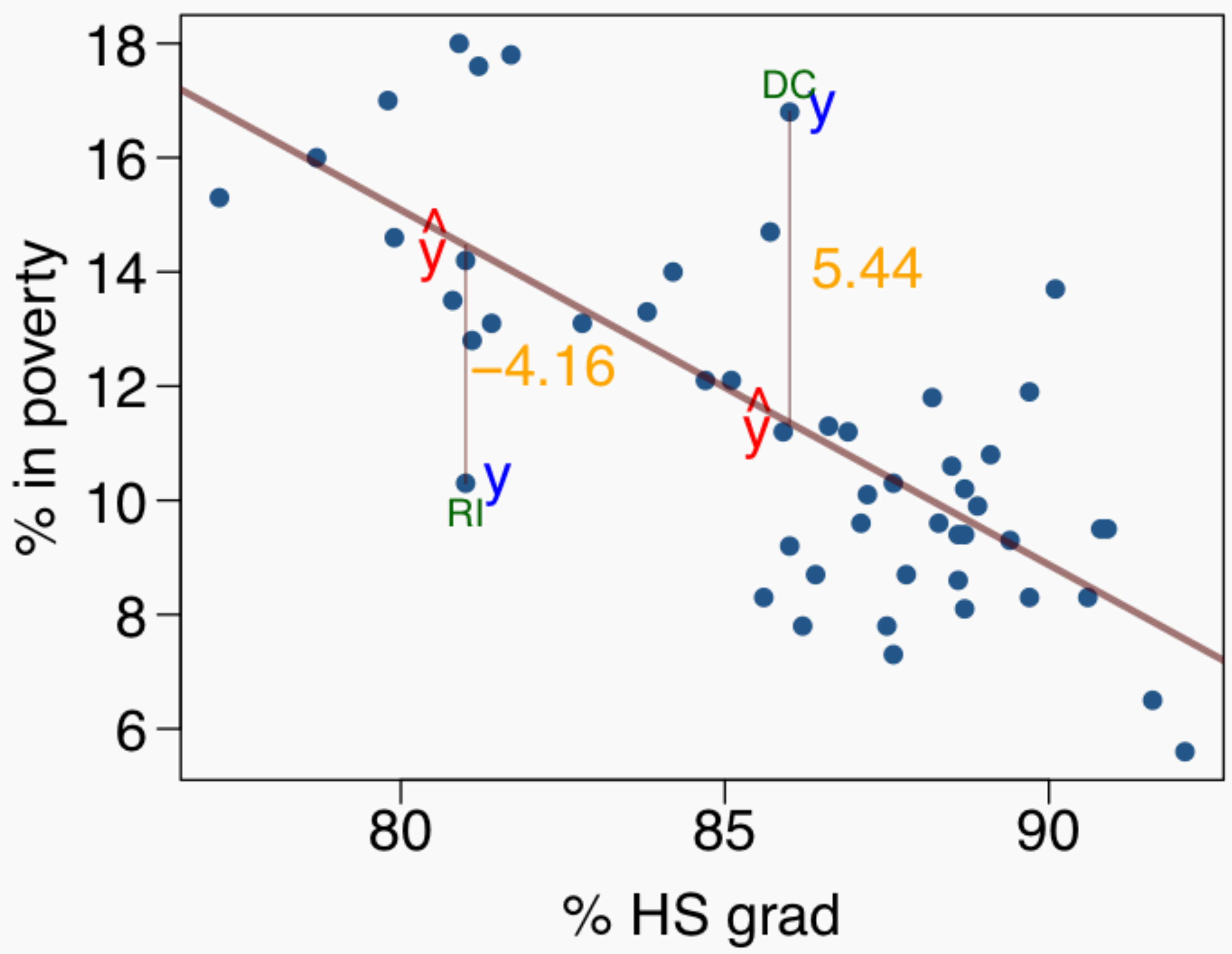



% living in poverty in DC is *5.44% more* than predicted.

# Residual Examples

We can think about a residual being the difference between our observed outcome ($y_i$) minus our predicted outcome.

$$e_i = y_i - \hat{y}_i = y_i - \beta_0 - \beta_1 x_i$$




% living in poverty in DC is *5.44% more* than predicted.

% living in poverty in RI is *4.16% less* than predicted.

# A measure for the best line

- We want a line that has small residuals - any idea what criteria we should use?

# A measure for the best line

- We want a line that has small residuals - any idea what criteria we should use?
  - Minimize the sum of squared residuals – *least squares*

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

# A measure for the best line

- We want a line that has small residuals - any idea what criteria we should use?
  - Minimize the sum of squared residuals – *least squares*

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

- Why least squares?

# A measure for the best line

- We want a line that has small residuals - any idea what criteria we should use?
    - Minimize the sum of squared residuals – *least squares*

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

- Why least squares?
    1. Most commonly used

# A measure for the best line

- We want a line that has small residuals - any idea what criteria we should use?
  - Minimize the sum of squared residuals – *least squares*

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

- Why least squares?
  1. Most commonly used
  2. Square is a nicer function than absolute value

# A measure for the best line

- We want a line that has small residuals - any idea what criteria we should use?
  - Minimize the sum of squared residuals – *least squares*

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

- Why least squares?
  1. Most commonly used
  2. Square is a nicer function than absolute value
  3. In many applications, a residual twice as large as another is more than twice as bad

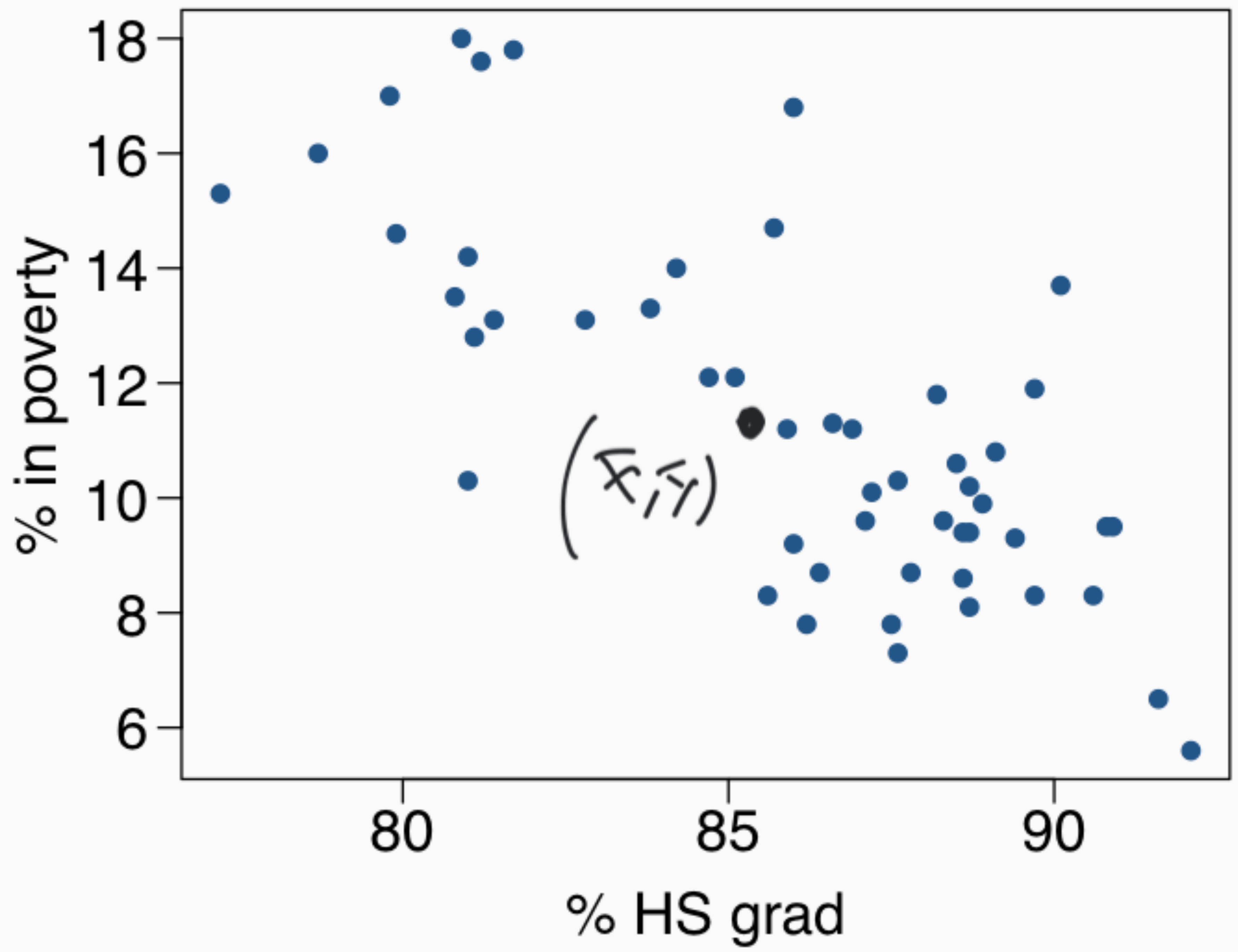
# The least squares line

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

Notation:

- Intercept:
  - Parameter: $\beta_0$
  - Point estimate: $b_0$
- Slope:
  - Parameter: $\beta_1$
  - Point estimate: $b_1$

# Data / Sample Statistics




| | % HS grad ($x$) | % in poverty ($y$) |
|---|---|---|
| mean | $\bar{x} = 86.01$ | $\bar{y} = 11.35$ |
| sd | $s_x = 3.73$ | $s_y = 3.1$ |
| | correlation | $R = -0.75$ |

What values of $b_0$ and $b_1$ will minimize the sum of squared residuals?

$$\underset{b_0, b_1}{\operatorname{argmin}} \sum_{i=1}^{n} \epsilon_i^2 = \underset{b_0, b_1}{\operatorname{argmin}} \sum_{i=1}^{2} (y_i - b_0 - b_1 x_i)^2$$

# Slope

The slope of the bivariate least squares regression line is given by

$$\beta_1 = \frac{Cov(X, Y)}{Var(X)} = \frac{\sigma_x \sigma_y}{\sigma_x^2} Cor(X, Y) = \frac{\sigma_y}{\sigma_x} \rho$$

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

# Slope

The slope of the bivariate least squares regression line is given by

$$\beta_1 = \frac{Cov(X,Y)}{Var(X)} = \frac{\sigma_X \sigma_y}{\sigma_X^2} Cor(X,Y) = \frac{\sigma_y}{\sigma_X}\rho$$

$$b_1 = \frac{s_y}{s_X} R$$

*In context:*

$$b_1 = \frac{3.1}{3.73} \times -0.75 = -0.62$$

# Slope

The slope of the bivariate least squares regression line is given by

$$\beta_1 = \frac{Cov(X,Y)}{Var(X)} = \frac{\sigma_X \sigma_Y}{\sigma_X^2} Cor(X,Y) = \frac{\sigma_Y}{\sigma_X} \rho$$

$$b_1 = \frac{s_Y}{s_X} R$$

*In context:*

$$b_1 = \frac{3.1}{3.73} \times -0.75 = -0.62$$

*Interpretation:*

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

# Intercept

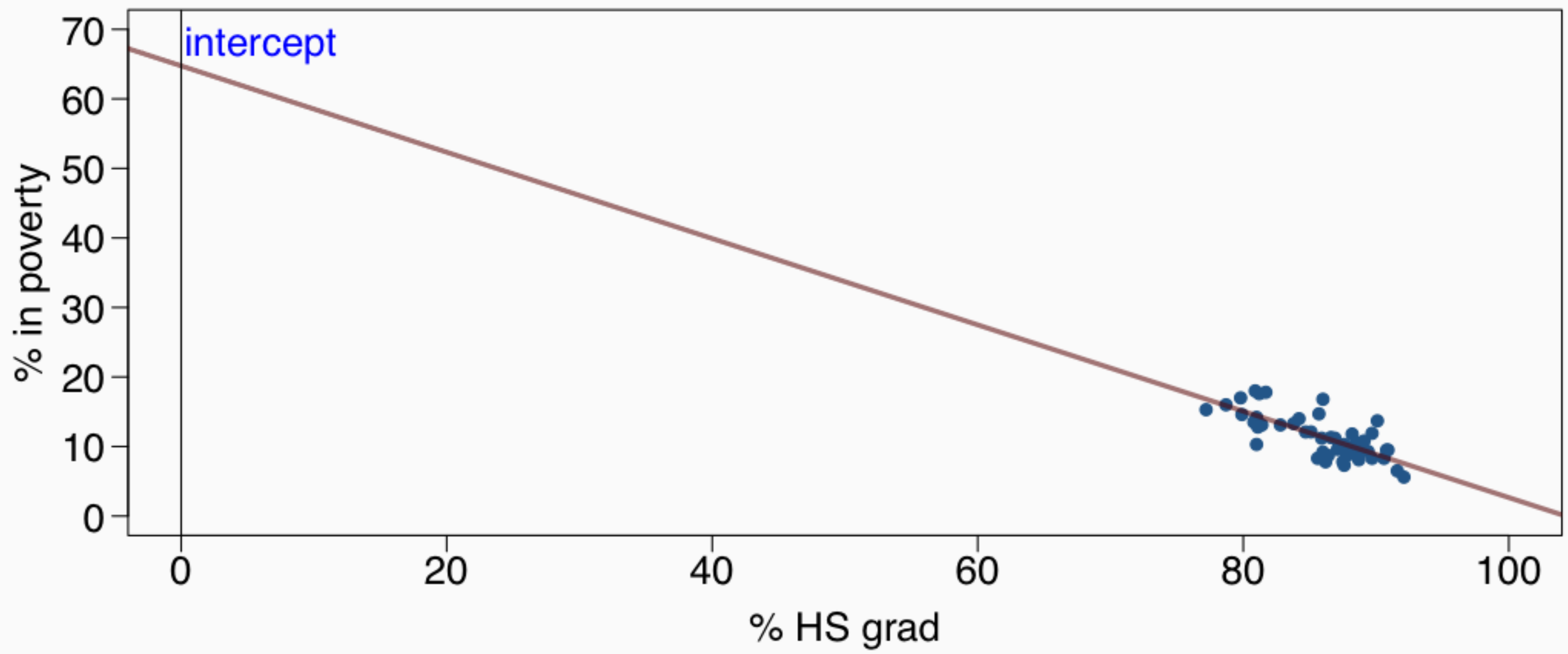
The intercept is where the line intersects the $y$-axis. To calculate the intercept for the least squares line we use the fact that the regression line *will always* pass through $(\bar{x}, \bar{y})$.

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

# Intercept

The intercept is where the line intersects the $y$-axis. To calculate the intercept for the least squares line we use the fact that the regression line *will always* pass through $(\bar{x}, \bar{y})$.

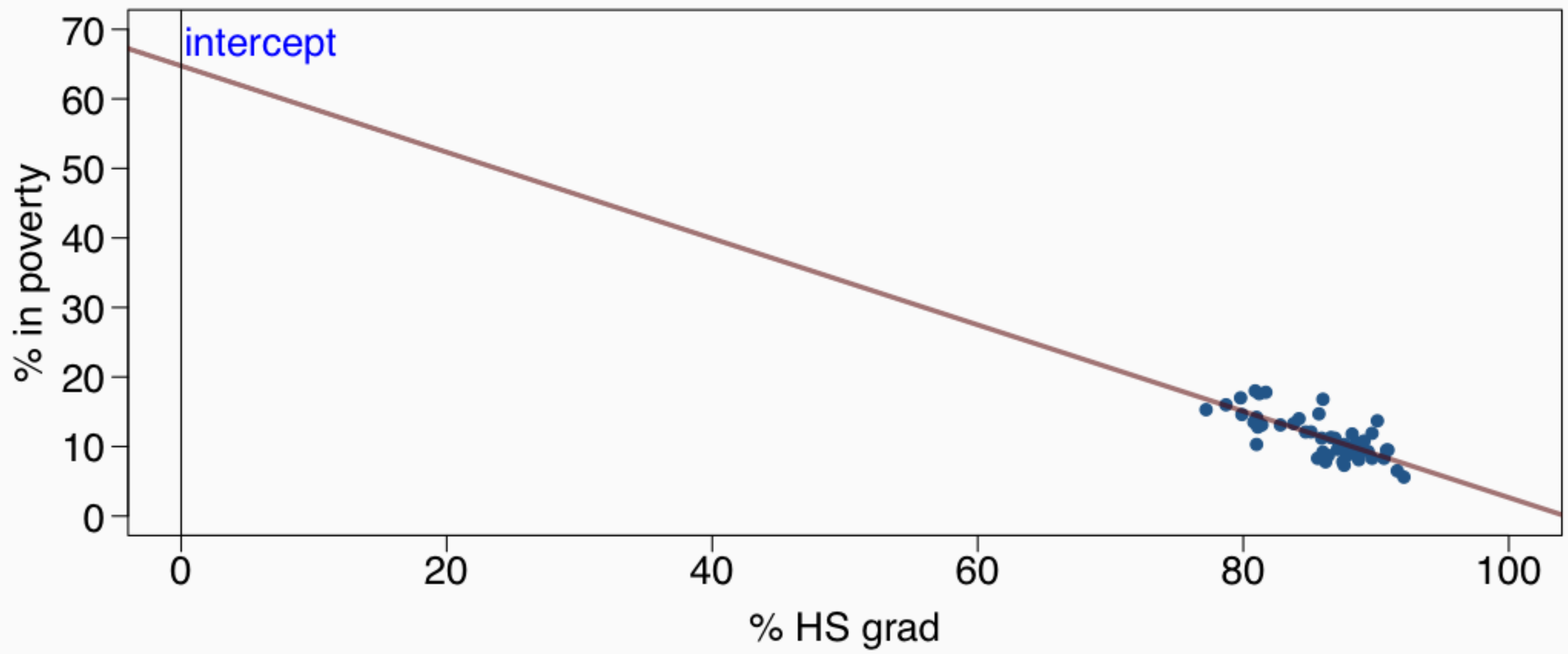
$$b_0 = \bar{y} - b_1\bar{x}$$



# Intercept

The intercept is where the line intersects the $y$-axis. To calculate the intercept for the least squares line we use the fact that the regression line *will always* pass through $(\bar{x}, \bar{y})$.

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




*In context:*

$$b_0 = 11.35 - (-0.62) \times 86.01 = 64.68$$

# Interpreting Intercepts

Which of the following is the correct interpretation of the intercept?

(a) For each % point increase in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

(b) For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.

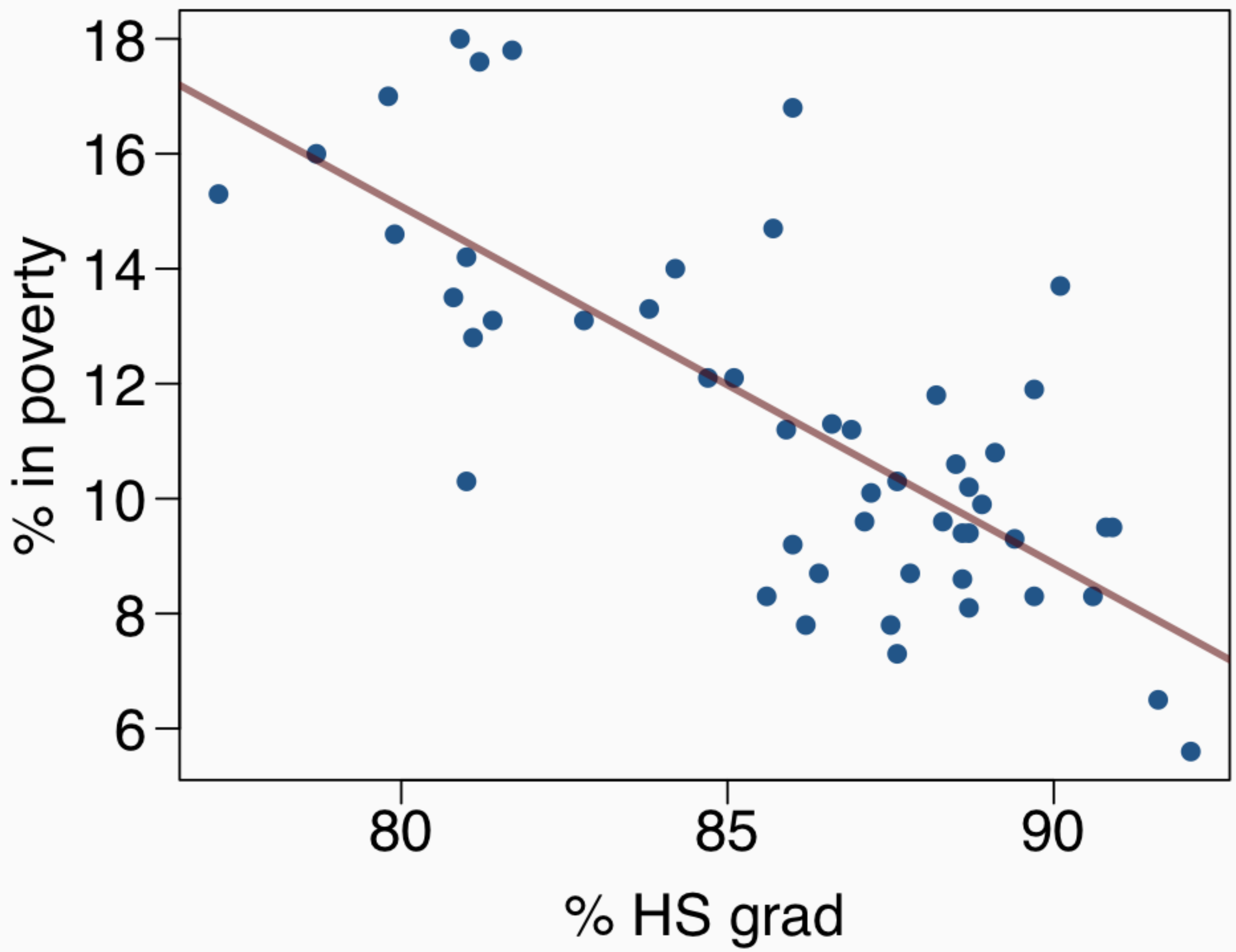
(c) Having no HS graduates leads to 64.68% of residents living below the poverty line. Causal statement

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

(e) In states with no HS graduates % living in poverty is expected to increase on average by 64.68%.

## Regression line

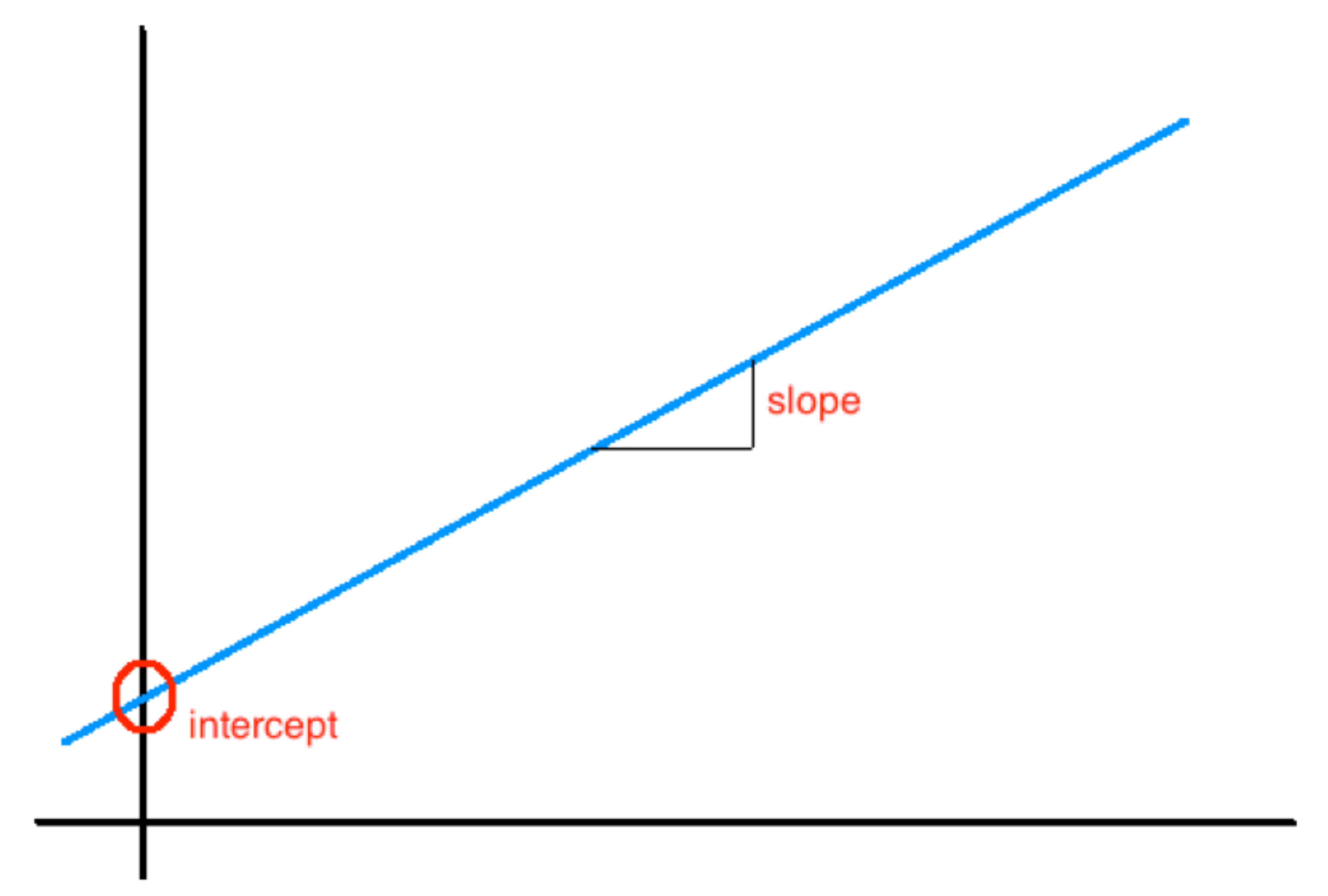
$$\widehat{[\% \; in \; poverty]} = 64.68 - 0.62 \; [\% \; HS \; grad]$$



# Interpretation of slope and intercept

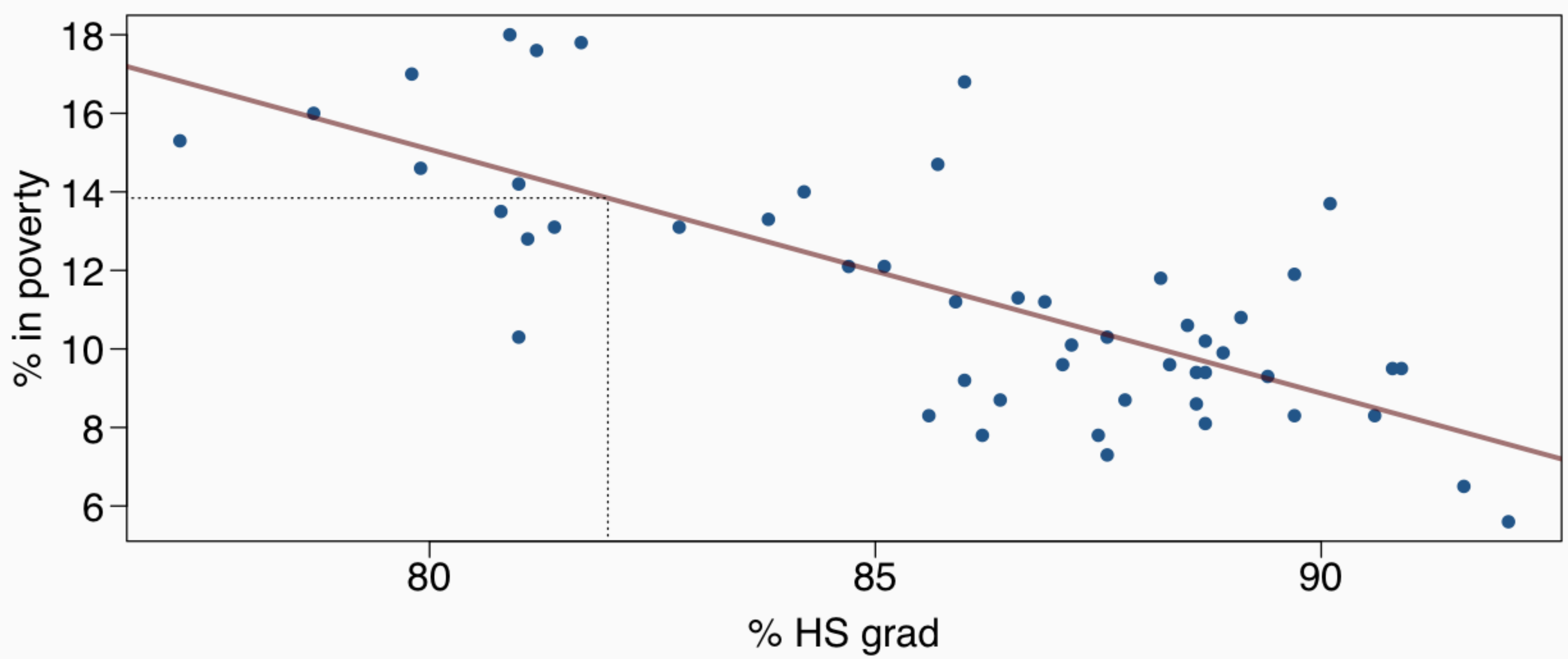
- *Intercept:* When $x = 0$, $y$ is expected to equal *the intercept* on average.

- *Slope:* For each *unit* increase in $x$, $y$ is expected to *increase/decrease* on average by *the slope*.



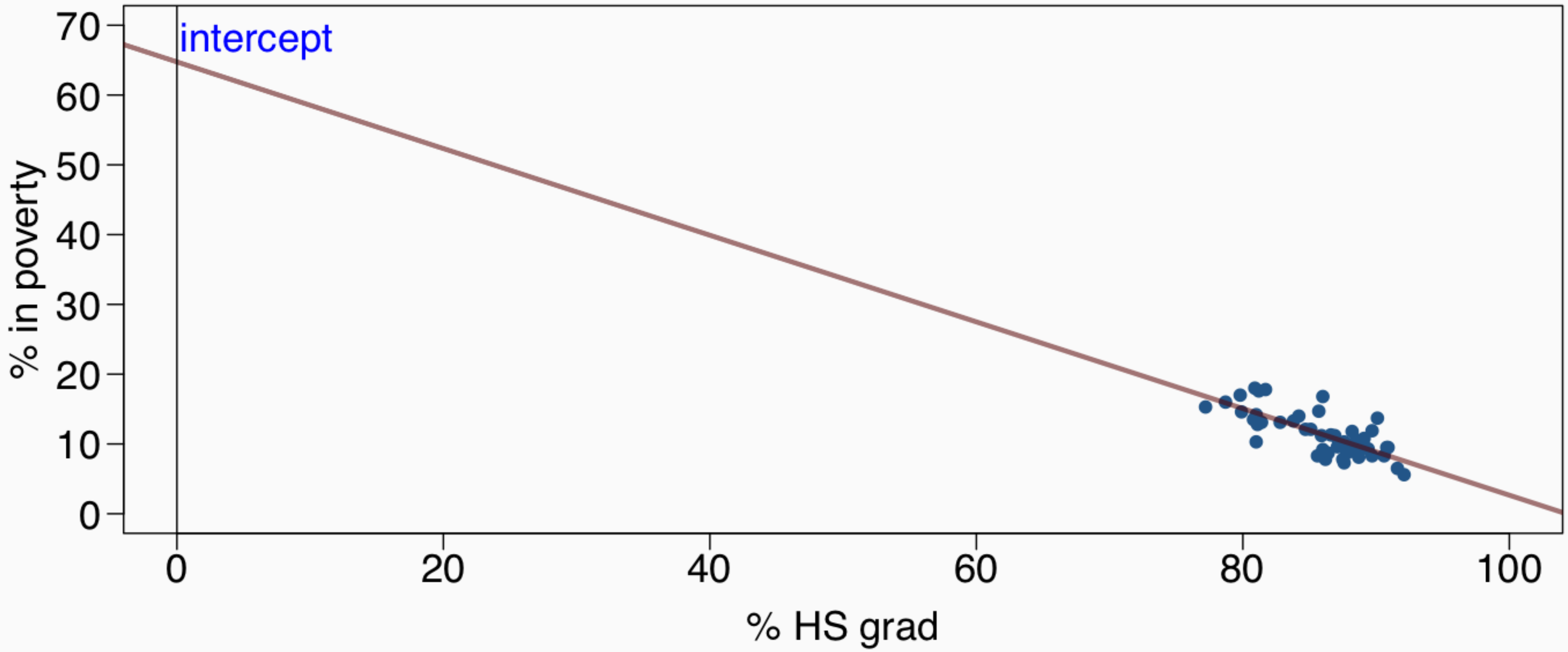
# Prediction

- Using the linear model we are able to predict the value of the response variable at any arbitrary value of the predictor variable by plugging in the value of $x$ in the linear model equation.
- There will be some uncertainty associated with the predicted value - we'll talk more about this next time.

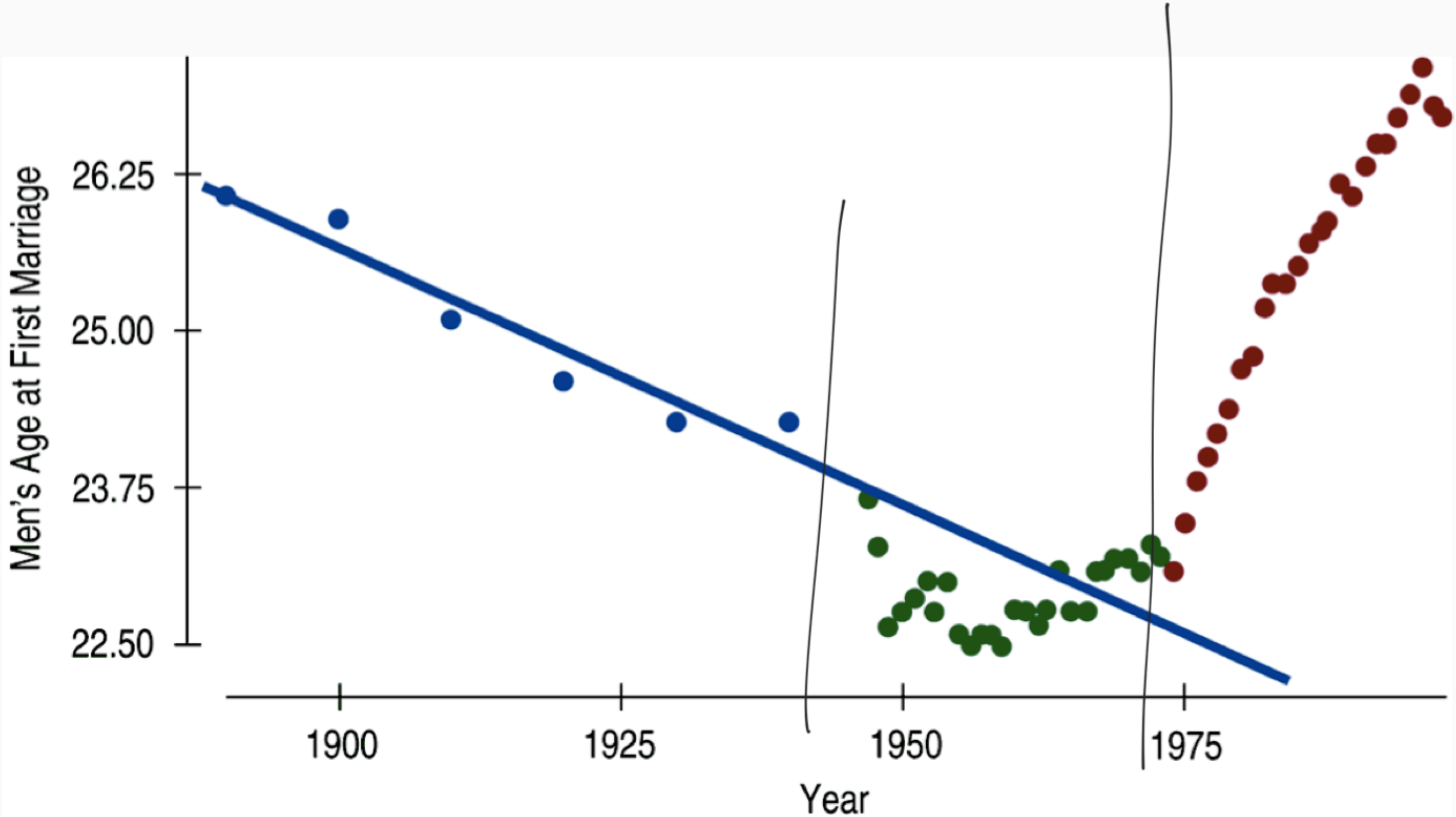


# Extrapolation

- Applying a model estimate to values outside of the range of the original data is called *extrapolation*.
- Sometimes the intercept might be an extrapolation.



# Examples of extrapolation



# Examples of extrapolation

BBC NEWS

Watch One-Minute World News

News Front Page
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
England
Northern Ireland
Scotland
Wales
UK Politics
Education
Magazine
Business
Health
Science & Environment
Technology
Entertainment
Also in the news

Last Updated: Thursday, 30 September, 2004, 04:04 GMT 05:04 UK

E-mail this to a friend    Printable version

## Women 'may outsprint men by 2156'

**Women sprinters may be outrunning men in the 2156 Olympics if they continue to close the gap at the rate they are doing, according to scientists.**

Women are set to become the dominant sprinters

An Oxford University study found that women are running faster than they have ever done over 100m.

At their current rate of improvement, they should overtake men within 150 years, said Dr Andrew Tatem.

The study, comparing winning times for the Olympic 100m since 1900, is published in the journal Nature.

However, former British Olympic sprinter Derek Redmond told the BBC: "I find it difficult to believe.

"I can see the gap closing between men and women but I can't necessarily see it being overtaken because mens' times are also going to improve."

# Examples of extrapolation

## Momentous sprint at the 2156 Olympics?

Women sprinters are closing the gap on men and may one day overtake them.

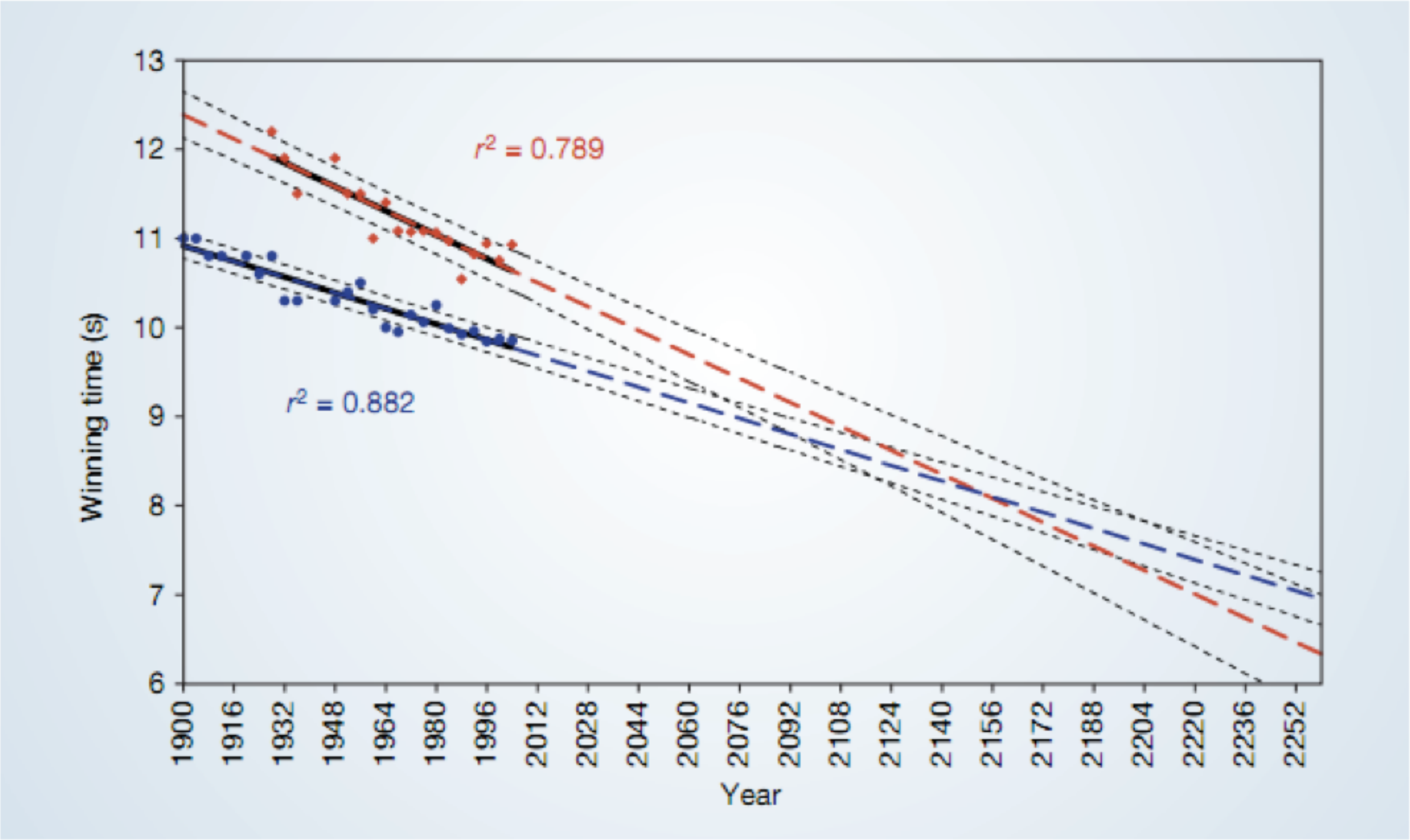
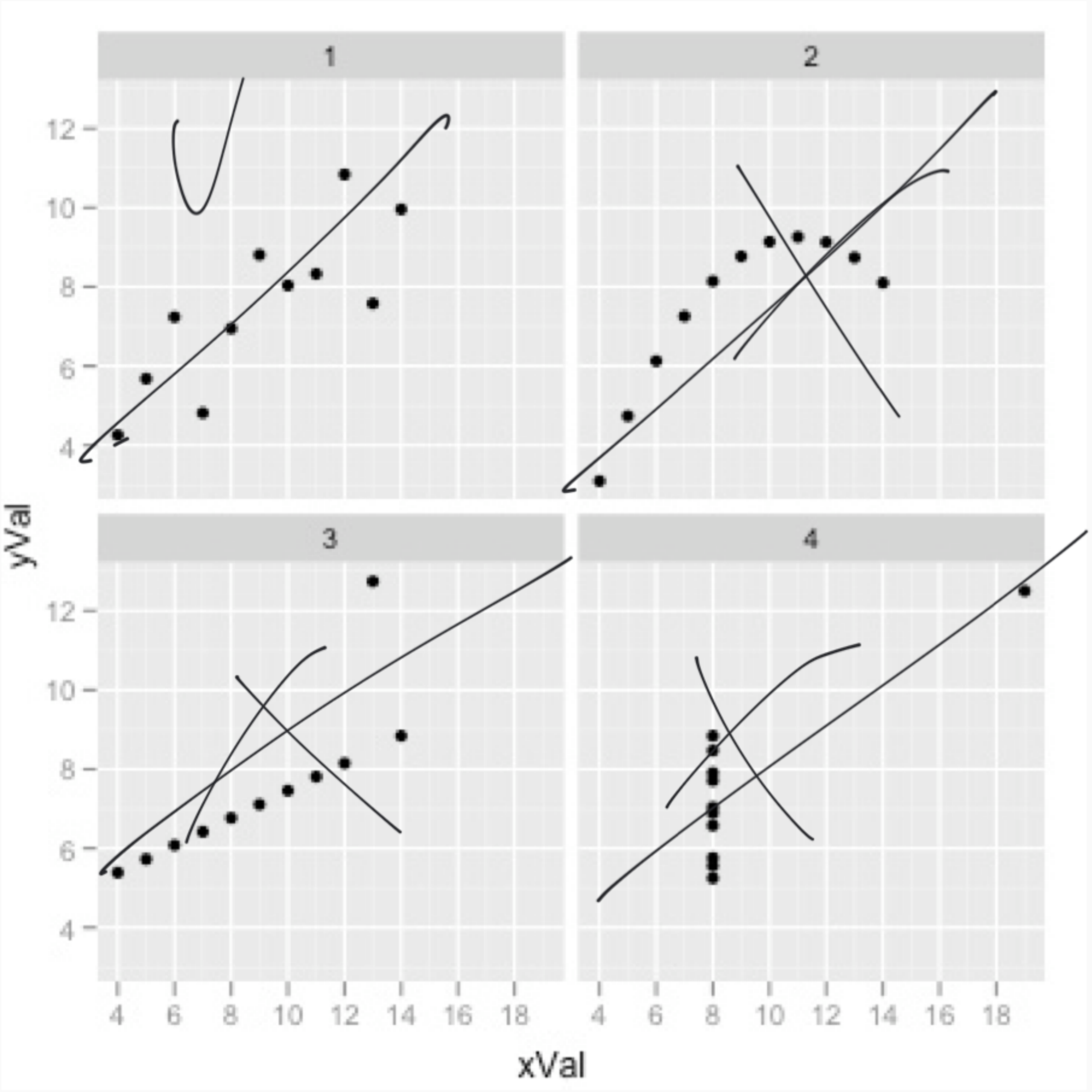



**Figure 1** The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) and 95% confidence intervals (dotted black lines) based on the available points are superimposed. The projections intersect just before the 2156 Olympics, when the winning women's 100-metre sprint time of 8.079 s will be faster than the men's at 8.098 s.

# Anscombe's Quartet



# Anscombe's Quartet - Data

| x1 | y1 | x2 | y2 | x3 | y3 | x4 | y4 |
|---|---|---|---|---|---|---|---|
| 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 14 | 9.96 | 14 | 8.10 | 14 | 8.84 | 8 | 7.04 |
| 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 4 | 4.26 | 4 | 3.10 | 4 | 5.39 | 19 | 12.50 |
| 12 | 0.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |

# Anscombe's Quartet - Data

| x1 | y1 | x2 | y2 | x3 | y3 | x4 | y4 |
|---|---|---|---|---|---|---|---|
| 10 | 8.04 | 10 | 9.14 | 10 | 7.46 | 8 | 6.58 |
| 8 | 6.95 | 8 | 8.14 | 8 | 6.77 | 8 | 5.76 |
| 13 | 7.58 | 13 | 8.74 | 13 | 12.74 | 8 | 7.71 |
| 9 | 8.81 | 9 | 8.77 | 9 | 7.11 | 8 | 8.84 |
| 11 | 8.33 | 11 | 9.26 | 11 | 7.81 | 8 | 8.47 |
| 14 | 9.96 | 14 | 8.10 | 14 | 8.84 | 8 | 7.04 |
| 6 | 7.24 | 6 | 6.13 | 6 | 6.08 | 8 | 5.25 |
| 4 | 4.26 | 4 | 3.10 | 4 | 5.39 | 19 | 12.50 |
| 12 | 0.84 | 12 | 9.13 | 12 | 8.15 | 8 | 5.56 |
| 7 | 4.82 | 7 | 7.26 | 7 | 6.42 | 8 | 7.91 |
| 5 | 5.68 | 5 | 4.74 | 5 | 5.73 | 8 | 6.89 |

All four datasets have the same regression line:

$$\hat{y} = 3 + 0.5x$$

# $R^2$

The strength of the fit of a linear model is often evaluated using a value called $R^2$.

# $R^2$

The strength of the fit of a linear model is often evaluated using a value called $R^2$.

- $R^2$ is calculated by squaring the correlation coefficient.

# $R^2$

The strength of the fit of a linear model is often evaluated using a value called $R^2$.

- $R^2$ is calculated by squaring the correlation coefficient.
- It has a useful interpretation - specifcially the $R^2$ equals the percent of variability in the response variable ($y$) that is explained by the predictor variable ($x$).

# $R^2$

The strength of the fit of a linear model is often evaluated using a value called $R^2$.

- $R^2$ is calculated by squaring the correlation coefficient.
- It has a useful interpretation - specifcially the $R^2$ equals the percent of variability in the response variable ($y$) that is explained by the predictor variable ($x$).
- $1 - R^2$ is therefore the amount variability that is not “explained” by the model.

# $R^2$

The strength of the fit of a linear model is often evaluated using a value called $R^2$.

- $R^2$ is calculated by squaring the correlation coefficient.
- It has a useful interpretation - specifcially the $R^2$ equals the percent of variability in the response variable ($y$) that is explained by the predictor variable ($x$).
- $1 - R^2$ is therefore the amount variability that is not "explained" by the model.
- Sometimes referred to as the coefficient of determination.

# $R^2$

The strength of the fit of a linear model is often evaluated using a value called $R^2$.

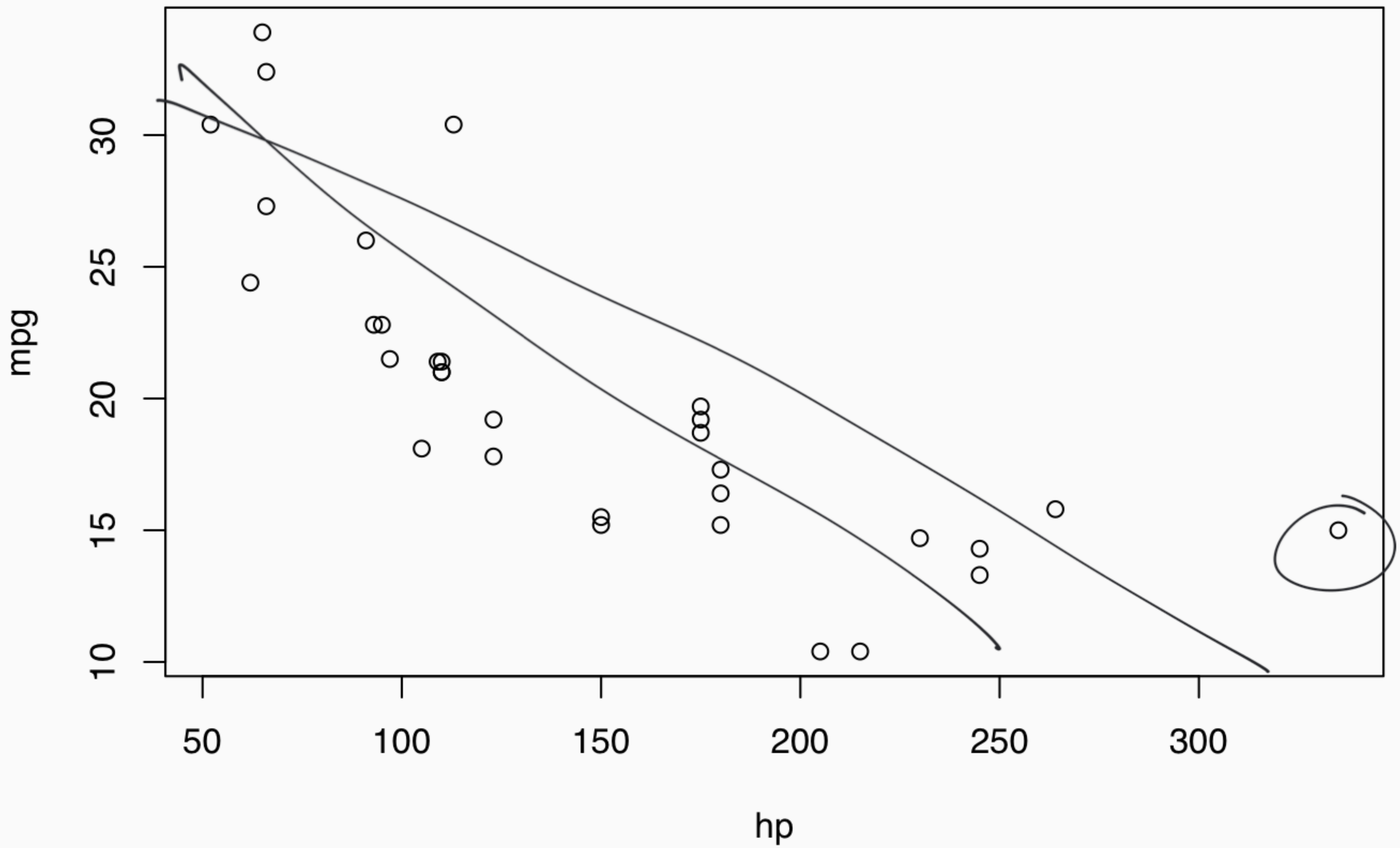
- $R^2$ is calculated by squaring the correlation coefficient.
- It has a useful interpretation - specifcially the $R^2$ equals the percent of variability in the response variable ($y$) that is explained by the predictor variable ($x$).
- $1 - R^2$ is therefore the amount variability that is not "explained" by the model.
- Sometimes referred to as the coefficient of determination.
- For the model we've been working with,

$$R^2 = (-0.75)^2 = 0.5625$$

# Modeling numerical variables

# mtcars

Data set from Motor Trend for 1973-74 model year cars.



# Least Squares fit

Find the least squares line that best describes these data.

| | mpg | hp |
|---|---|---|
| mean | 20.09 | 146.69 |
| sd | 6.03 | 68.56 |

$n = 32 \qquad R = -0.776$

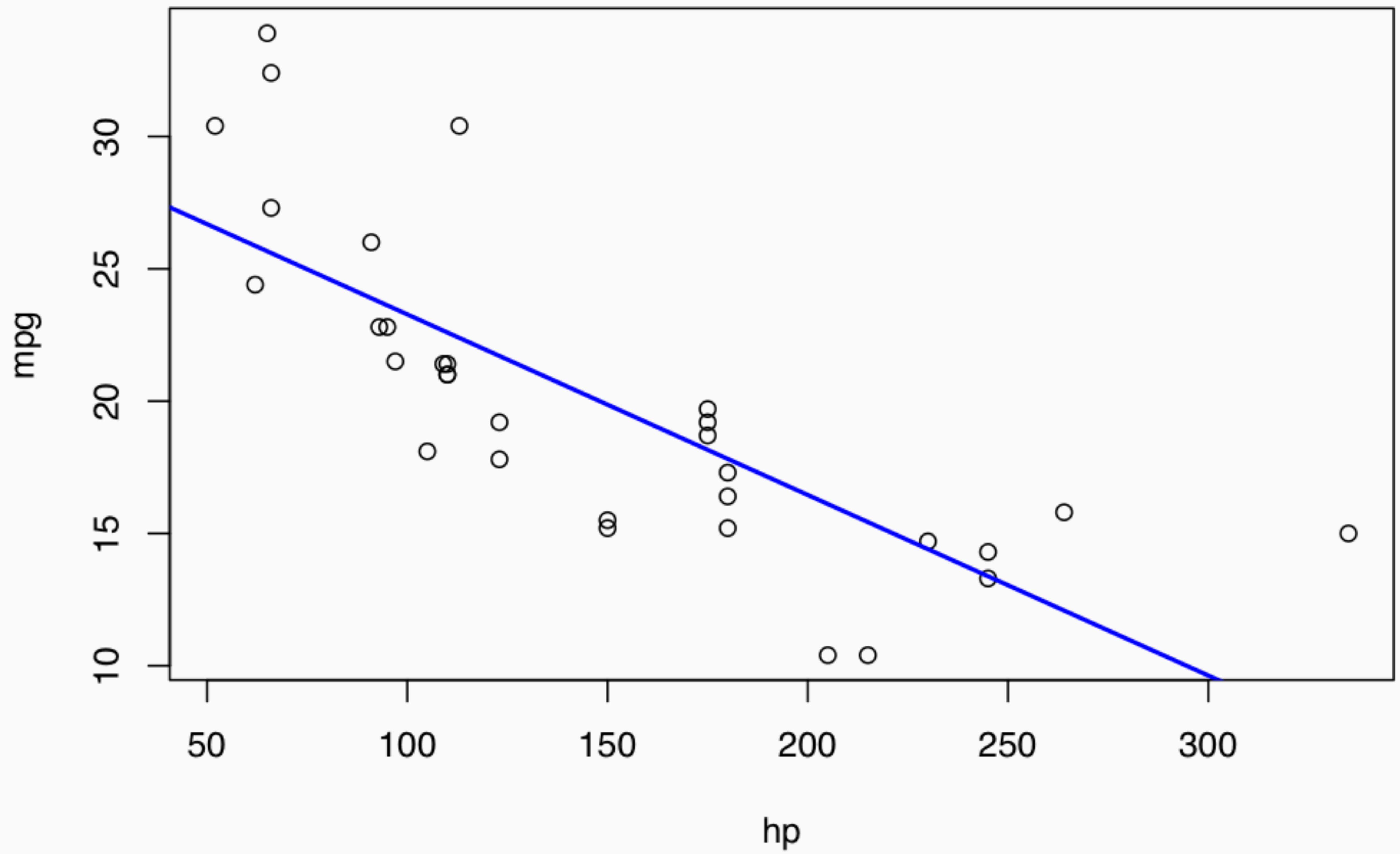
$$b_1 = \frac{s_y}{s_x} R = \frac{6.03}{68.58} \cdot (-0.776) = -0.068$$

$$\bar{y} = b_0 + b_1 \bar{x} \Rightarrow b_0 = \bar{y} - b_1 \bar{x}$$

$$= 20.09 - (-0.0068)(146.69)$$

$$\approx 30$$

# mtcars - line




| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 30.0989 | 1.6339 | 18.42 | 0.0000 |
| hp | -0.0682 | 0.0101 | -6.74 | 0.0000 |