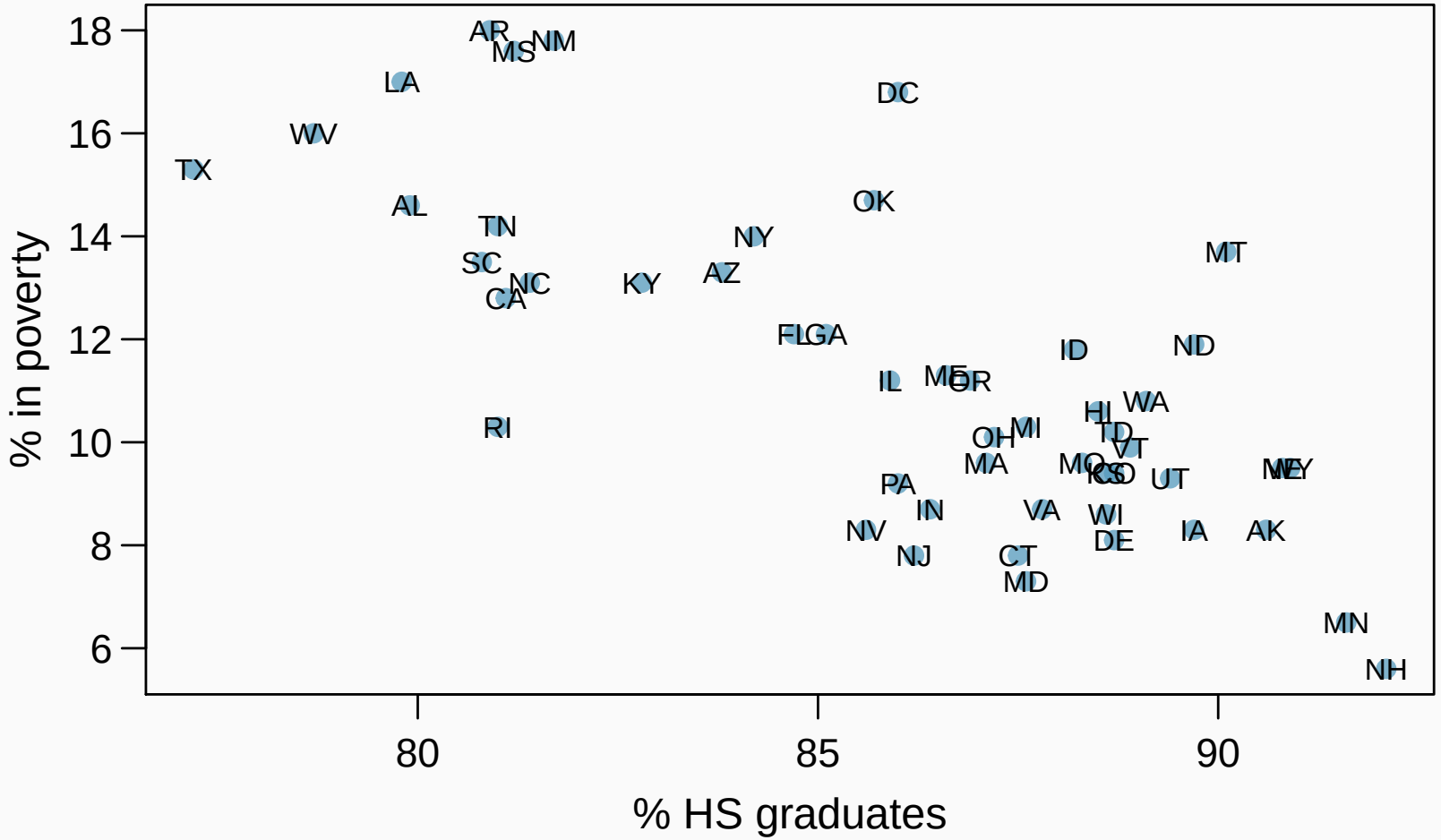
STAT 220 Lecture Slides
Least Square Regression Line

Yibi Huang
Department of Statistics
University of Chicago

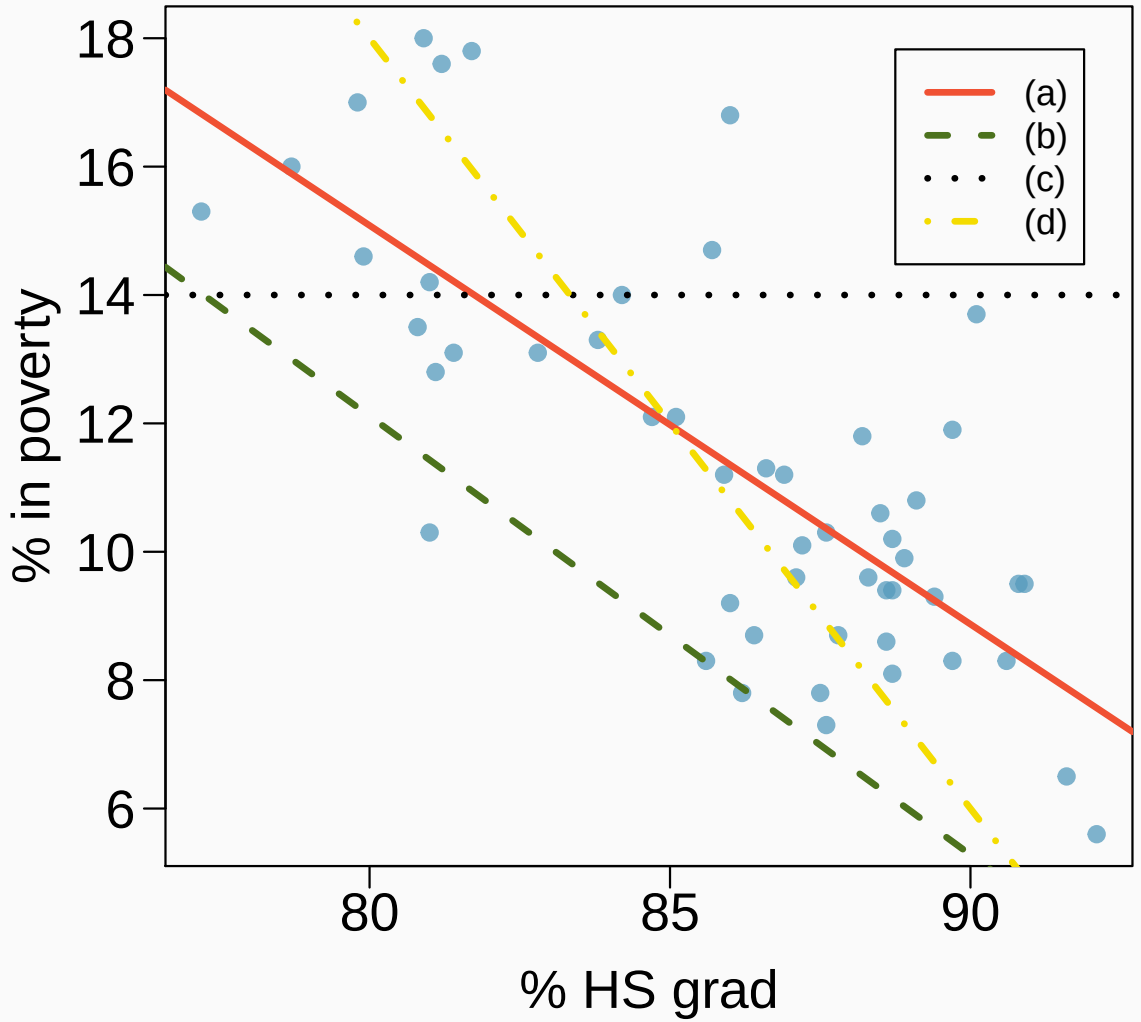
Poverty vs. HS graduate rate

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).



Eyeballing the line

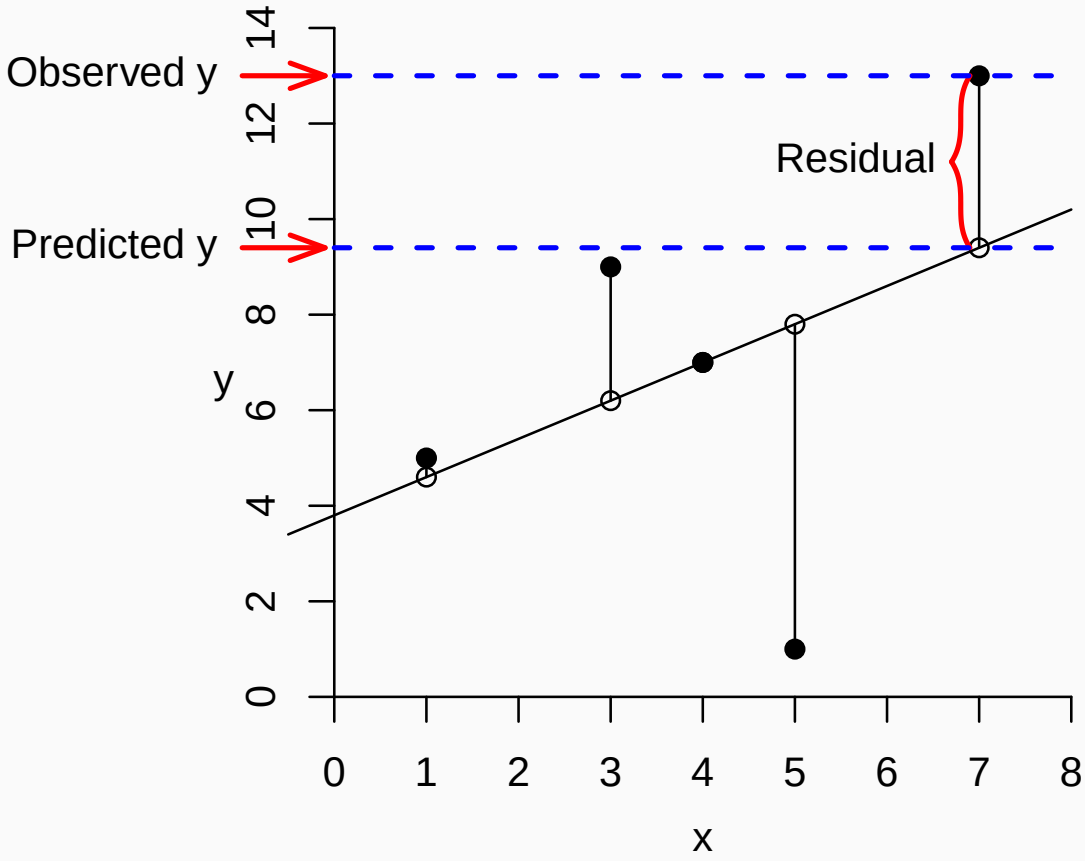
Which line appears to best fit the linear relationship between % in poverty and % HS grad?



Residuals (Prediction Errors)

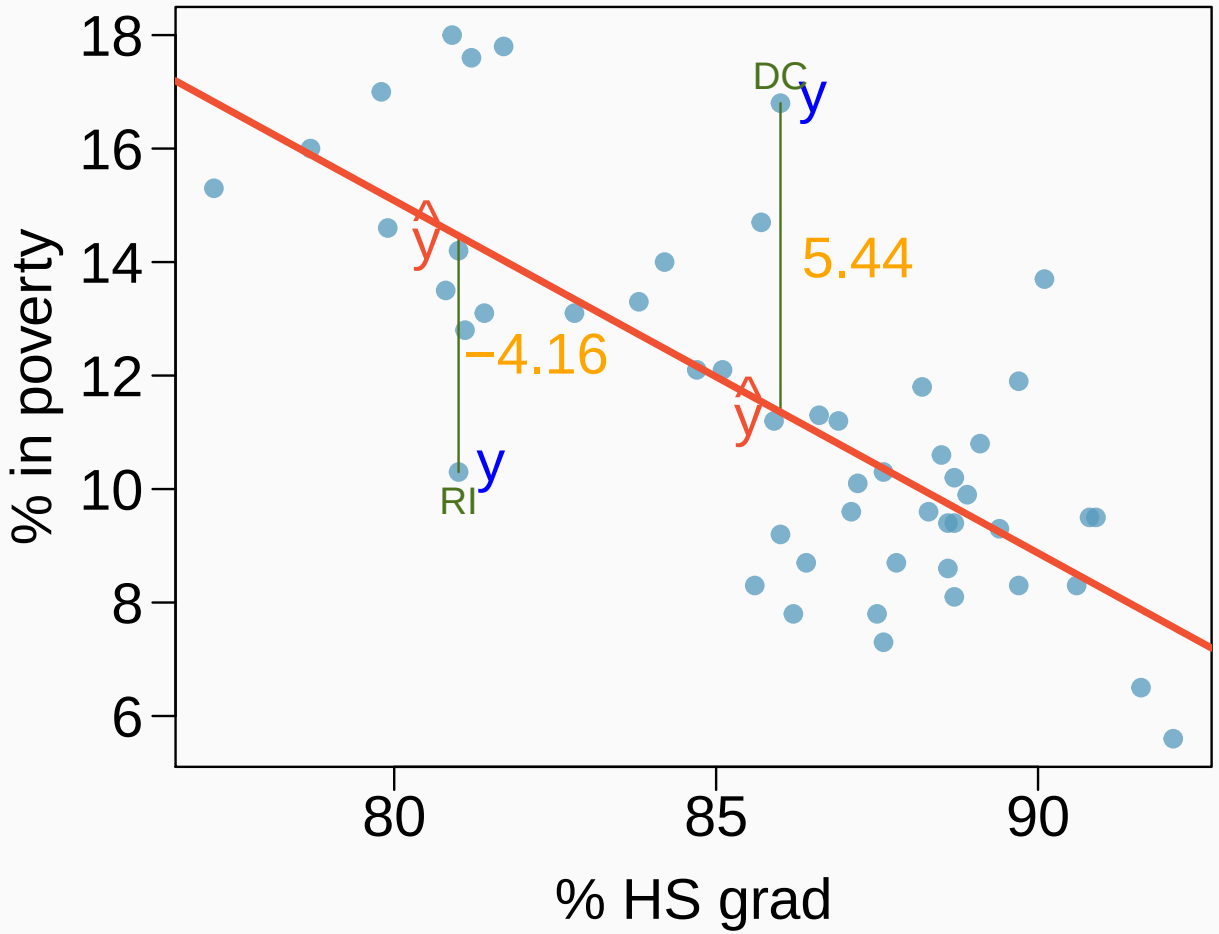
The *residual* (e_i) of the ith observation (x_i, y_i) is

$$\underset{\text{(Residual)}}{e_i} = \underset{\text{(Observed } y\text{)}}{y_i} - \underset{\text{(Predicted } y\text{)}}{\hat{y}_i}$$



- *Residuals* are the (signed) *vertical* distances from data points to model line, not the *shortest* distances
- Points above/below the model line have positive/negative residuals.

Residuals (cont.)



- % living in poverty in DC is 5.44% more than predicted.
- % living in poverty in RI is 4.16% less than predicted.

Criteria for Choosing the "Best" Line

We want a line $y = a + bx$ having small residuals:

Option 1: Minimize the sum of absolute values of residuals

$$|e_1| + |e_2| + \cdots + |e_n| = \sum_i |y_i - \widehat{y_i}| = \sum_i |y_i - a - bx_i|$$

- Difficult to compute. Nowadays possible by computer technology
- Giving less penalty to points with large residuals.
  The line selected is less sensitive to outliers

Option 2: Minimize the sum of squared residuals – *least square method*

$$e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_i (y_i - \widehat{y_i})^2 = \sum_i (y_i - a - bx_i)^2$$

- Easier to compute by hand and using software
- Giving more penalty to points with large residuals.
  A residual 2x as large as another is often more than 2x as bad.
  The line selected is more sensitive to outliers

Equation of the Least-Square (LS) Regression Line

The *least-square (LS) regression line* is the line $y = a + bx$ that minimizes the sum of squared errors:

$$\sum_i e_i^2 = \sum_i (y_i - \widehat{y}_i)^2 = \sum_i (y_i - a - bx_i)^2$$

The slope and intercept of the LS regression line can be shown by math to be

$$b = \text{slope} = r \cdot \frac{\text{SD}_y}{\text{SD}_x}$$

$$a = \text{intercept} = \bar{y} - \text{slope} \cdot \bar{x}$$

Properties of the LS Regression Line

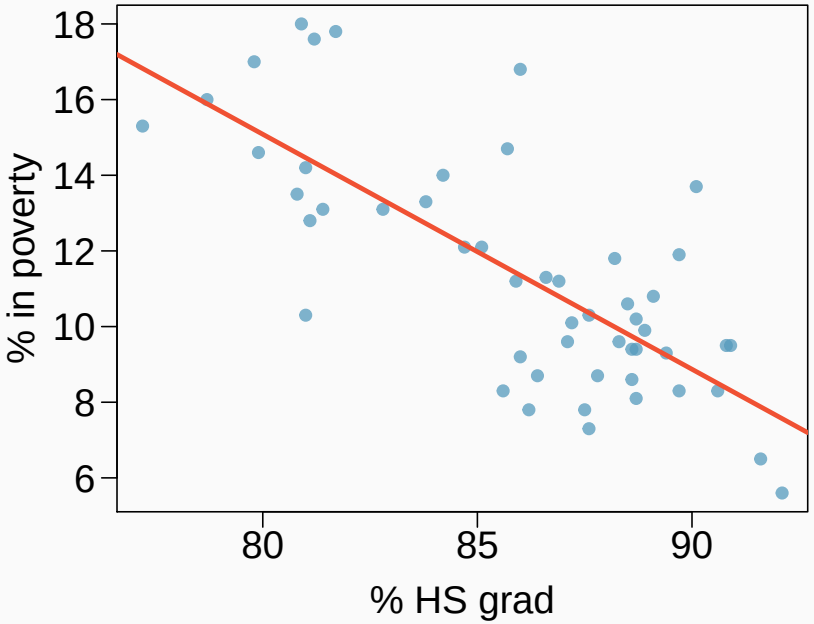
$$\widehat{y} = \underbrace{\text{intercept}}_{=\bar{y}-\text{slope}\cdot\bar{x}} + \text{slope} \cdot x$$

$$\Leftrightarrow \widehat{y} - \bar{y} = \text{slope} \cdot (x - \bar{x}) = r\frac{\text{SD}_y}{\text{SD}_x}(x - \bar{x})$$

$$\Leftrightarrow \underbrace{\frac{\widehat{y} - \bar{y}}{\text{SD}_y}}_{z\text{–score of } \widehat{y}} = r \cdot \underbrace{\frac{x - \bar{x}}{\text{SD}_x}}_{z\text{–score of } x}$$

- The LS regression line *always passes through* $(\bar{x}, \bar{y})$
- As x goes up by 1 SD of x, the predicted value $\widehat{y}$ only goes up by $r \times$ (SD of y)
- When $r = 0$, the LS regression line is horizontal $y = \bar{y}$, and the predicted value $\widehat{y}$ is *always the mean* $\bar{y}$

Poverty vs. HS graduate rate



	% HS grad (x)	% in poverty (y)
mean	$\bar{x} = 86.01$	$\bar{y} = 11.35$
sd	$SD_x = 3.73$	$SD_y = 3.1$
	correlation	$r = -0.75$

The *slope* and the *intercept* of the least square regression line is

$$\text{slope} = r\frac{SD_y}{SD_x} = (-0.75) \times \frac{3.1}{3.73} = -0.62$$

$$\text{intercept} = \bar{y} - \text{slope} \cdot \bar{x} = 11.35 - (-0.62) \times 86.01 = 64.68$$

So the equation of the least square regression line is

$$\widehat{\%\ \textit{in poverty}} = 64.68 - 0.62\ (\%\ \textit{HS grad})$$

Interpretation of Slope

The *slope* indicates how much the response changes *associated* with a unit change in x *on average* (may NOT be *causal*, unless the data are from an experiment).

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

- For each additional % point in HS graduation rate, we would expect the % in poverty to be lower on average by 0.62%.
- If a state manages to bring up its HS graduation rate by 1%, will its living-in-poverty rate lowers by 0.62%?
  Ans: Not necessarily, may not be causal.

Interpretation of the Intercept

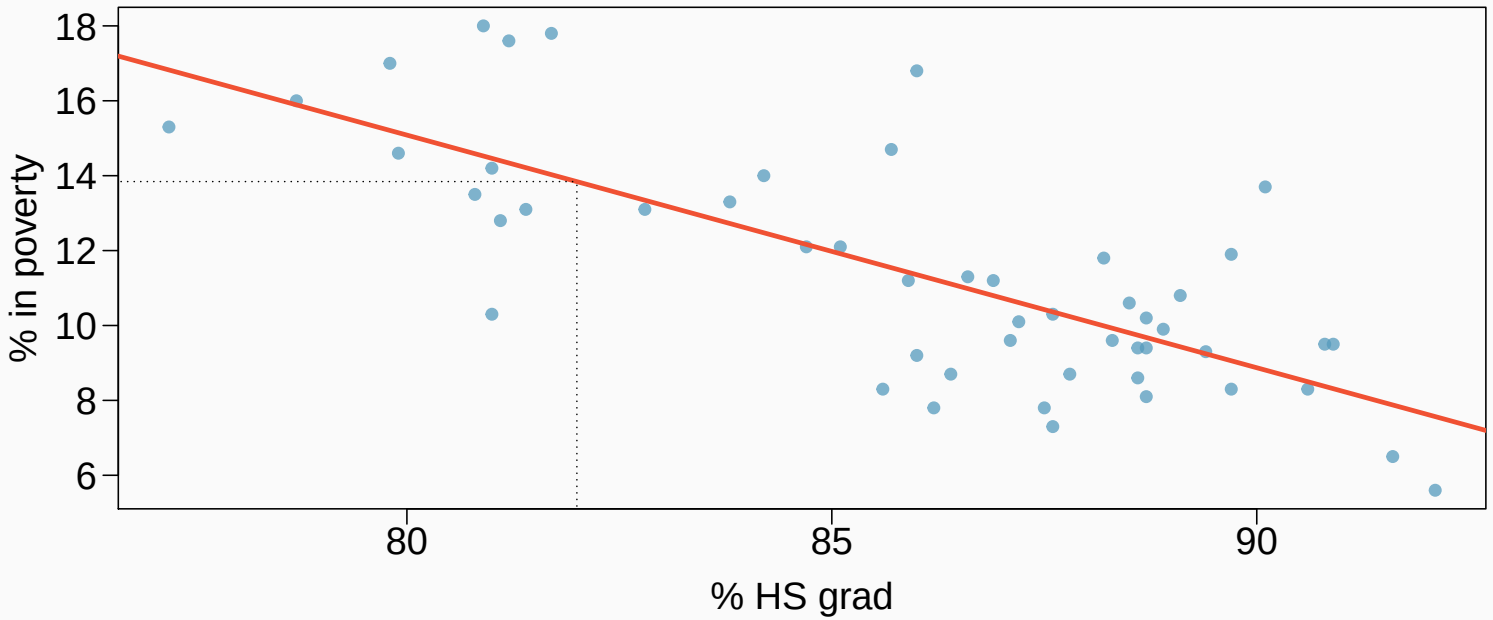
The *intercept* is the predicted value of response when $x = 0$, which might not have a practical meaning when $x = 0$ is not a possible value

$$\widehat{\%\ in\ poverty} = 64.68 - 0.62(\%\ HS\ grad)$$

- States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line
  - meaningless. There is no such state.

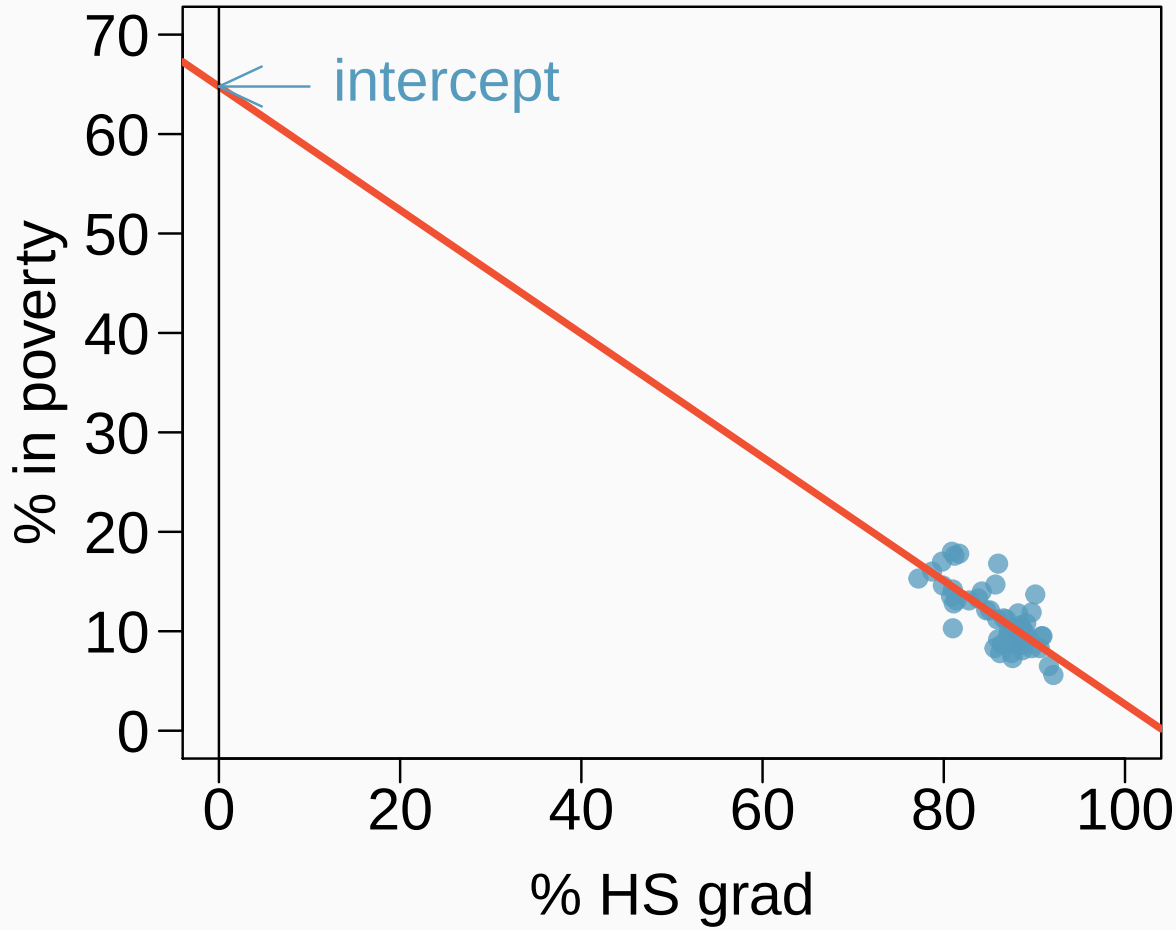
Prediction

- Using the linear model to predict the value of the response variable for a given value of the explanatory variable is called *prediction*, simply by plugging in the value of x in the linear model equation.
- There will be some uncertainty associated with the predicted value.



Extrapolation

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



Examples of Extrapolation

BBC NEWS

Watch One-Minute World News

News Front Page

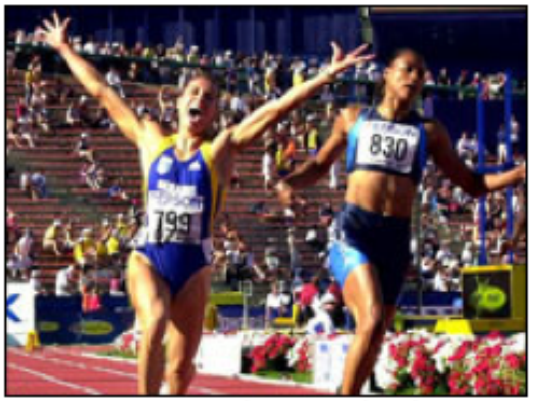
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
England
Northern Ireland
Scotland
Wales
UK Politics
Education
Magazine
Business
Health
Science & Environment
Technology
Entertainment
Also in the news

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.

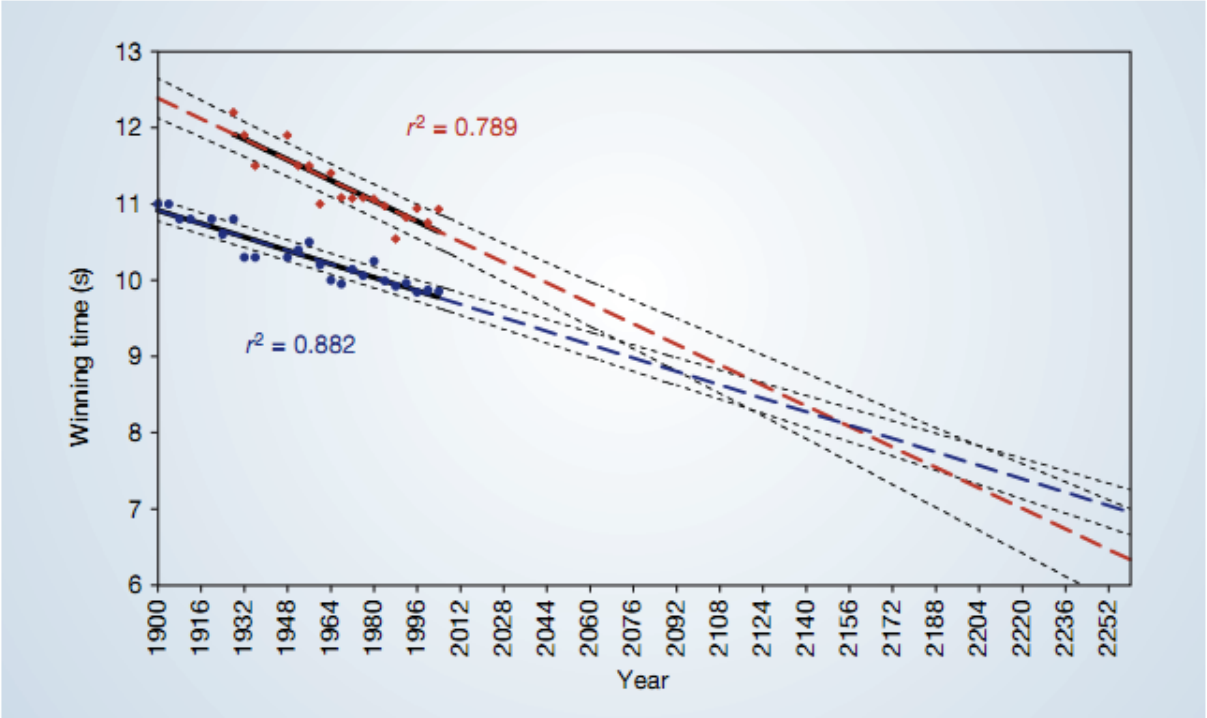


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively) 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.

There Are Two LS Regression Lines (1)

Recall the LS regression line for predicting poverty rate from HS graduation rate is

$$\widehat{\%\ in\ poverty} = 64.68 - 0.62\ (\%\ HS\ grad)$$

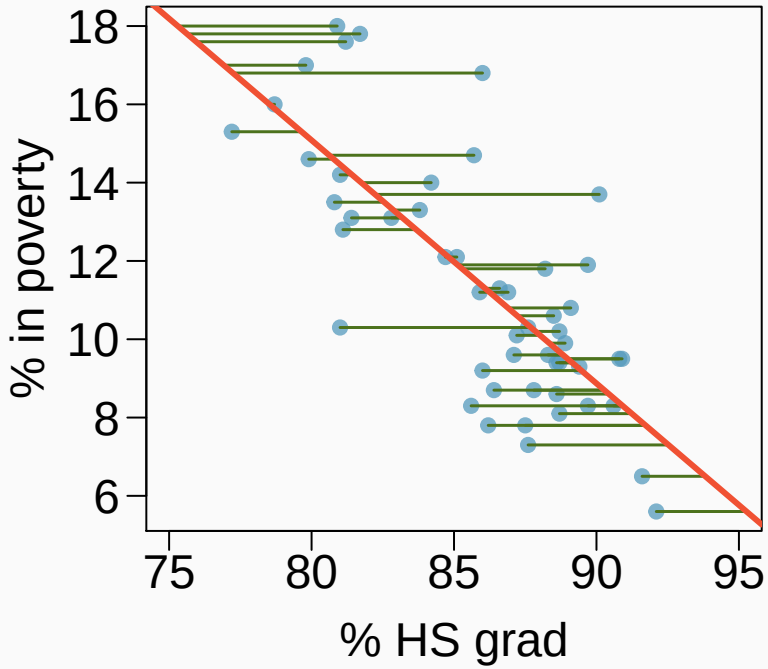
For a state with 80% HS graduation rate, the predicted poverty rate is

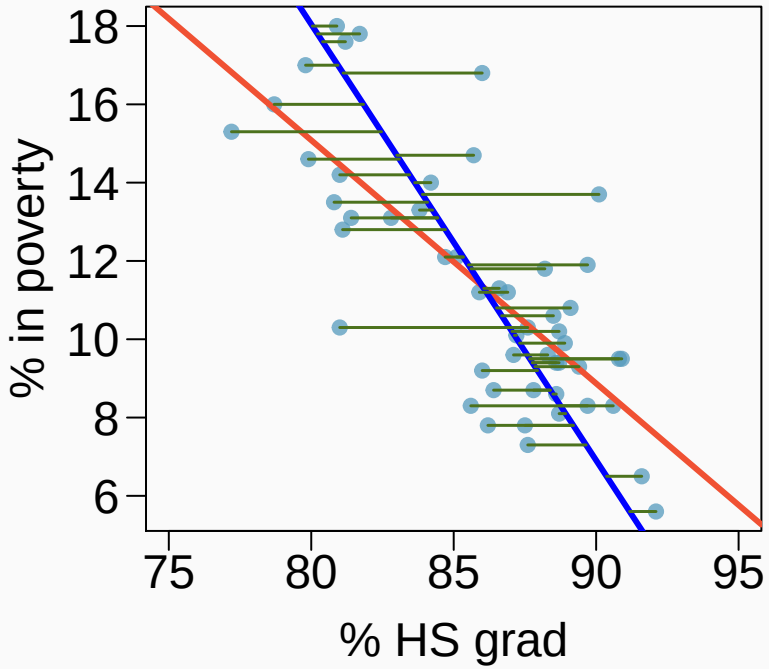
$$\widehat{\%\ in\ poverty} = 64.68 - 0.62 \times 80 = 15.08\%$$

For another state with 15.08% of residents living in poverty, will the predicted HS graduation rate to be 80%?

There Are Two LS Regression Lines (2)

The *prediction errors (residuals)* for predicting *X from Y* are the *horizontal* distances from data points to the line.



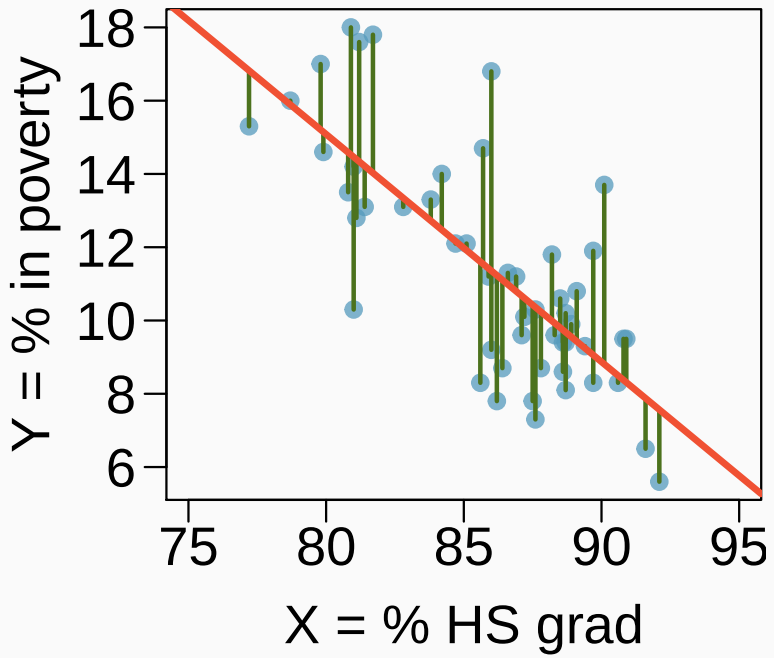


Which line, Red or Blue, has smaller prediction errors in general when predicting % HS grad from % in poverty?

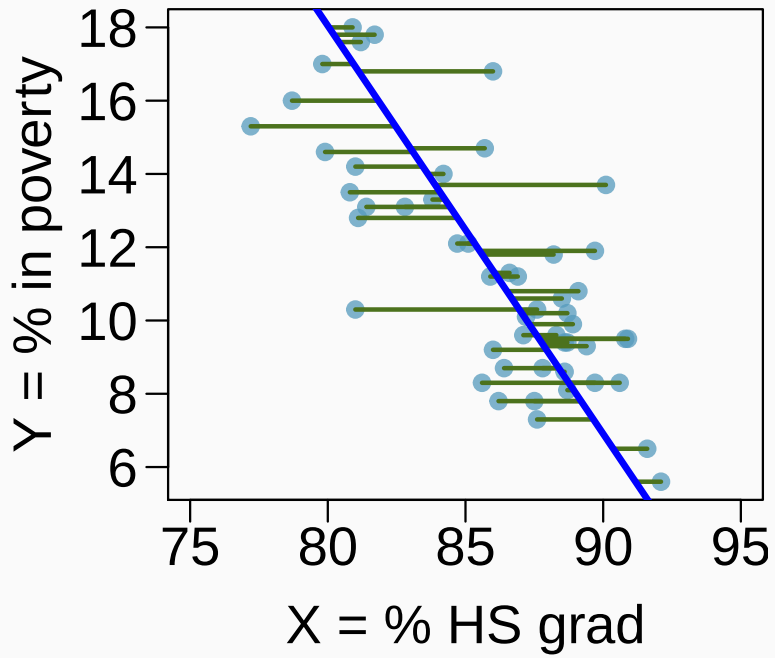
There Are Two LS Regression Lines (3)

The regression line for predicting $\begin{matrix} Y \text{ from } X \\ X \text{ from } Y \end{matrix}$ minimizes the sum of squared $\begin{matrix} \text{vertical} \\ \text{horizontal} \end{matrix}$ distances from the points to the line.

Predicting Y from X



Predicting X from Y



The two lines are different.

There Are Two LS Regression Lines (4)

The equation for predicting *% in poverty* from *% HS grad* is only for predicting *% in poverty* from *% HS grad*. One cannot plug in *% in poverty* to solve for the predicted *% HS grad*.

Recall that correlation does not distinguish between X and Y.
For regression, the two variables play different roles and are not interchangeable.

There Are Two LS Regression Lines (5)

	% HS grad (y)	% in poverty (x)
mean	86.01	11.35
sd	3.73	3.1
	correlation	$r = -0.75$

Find the equation for the LS regression line that predicts % HS grad from % in poverty.

- What is x? What is y? $x =$ % in poverty, $y =$ % HS grad
- slope $= r\,\mathrm{SD}_y/\mathrm{SD}_x = -0.75 \times 3.73/3.1 = -0.90$
- intercept $= \bar{y} - (\text{slope}) \cdot \bar{x} = 86.01 - (-0.90) \times 11.35 = 96.2$
- equation: $\widehat{\%\ \mathit{HS\ grad}} = 96.2 - 0.90\ (\%\ \mathit{in\ poverty})$

For a state with 15.08% living in poverty, the predicted HS graduation rate is $96.2 - 0.90 \times 15.08 = 82.628\%$, not 80%.

Properties of Residuals

Residuals for a least square regression line have the following properties.

1. Residuals always **sum to zero**, $\sum_{i=1}^{n} e_i = 0$.
    - If the sum > 0, can you improve the prediction?
2. Residuals and the explanatory variable x_i's have **zero correlation**.
    - If non-zero, the residuals can be predicted by x_i's, not the best prediction.
    - Residuals are the part in the response that CANNOT be explained or predicted linearly by the explanatory variables.

Variances of Residuals and Predicted Values

Recall

$$\underset{\text{(observed)}}{y_i} = \underset{\text{(predicted)}}{\widehat{y_i}} + \underset{\text{(residual)}}{e_i}$$

There is a nontrivial identity:

$$\underset{\text{(variance of } y\text{)}}{\frac{1}{n-1}\sum_i (y_i - \overline{y})^2} = \underset{\substack{\text{(variance of } \hat{y}\text{)} \\ \| \\ \text{(variability of } y \text{ that can be explained by } x\text{)}}}{\frac{1}{n-1}\sum_i (\hat{y}_i - \overline{y})^2} + \underset{\substack{\text{(variance of residuals)} \\ \| \\ \text{(variability of } y \text{ that cannot be explained by } x\text{)}}}{\frac{1}{n-1}\sum_i e_i^2}$$

R^2 = **R-squared** = r^2

Moreover, one can show that

$$r^2 = \frac{\sum_i (\hat{y}_i - \overline{y})^2}{\sum_i (y_i - \overline{y})^2} = \frac{\text{Variance of predicted } y\text{'s}}{\text{Variance of observed } y\text{'s}}$$

That is,

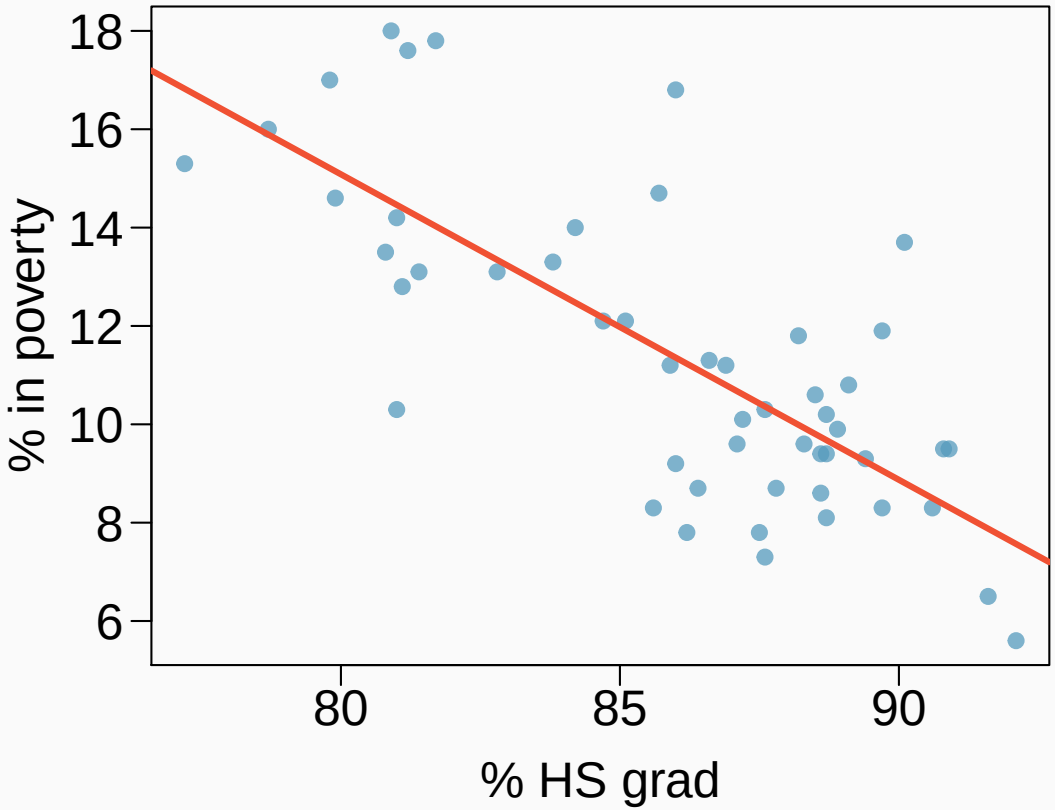
$$\begin{aligned} R^2 = r^2 &= \text{the square of the correlation coefficient} \\ &= \text{the proportion of variation in the response} \\ &\quad \text{explained by the explanatory variable} \end{aligned}$$

The remainder of the variability is explained by variables not included in the model or by inherent randomness in the data.

$$1 - r^2 = \frac{\sum_i e_i^2}{\sum_i (y_i - \overline{y})^2} = \frac{\text{Variance of Residuals}}{\text{Variance of } y}$$

Interpretation of R-squared

For the model we've been working with, $R^2 = r^2 = (-0.75)^2 \approx 0.56$, which means — 56% of the variability in the % of residents living in poverty among the 51 states is explained by the variable "% of HS grad".



Summary

- The residual of a data point is
  - the diff. between the observed y and the predicted y,
  - the signed vertical distance (not the shortest distance) from the data point to model line.
- The *least-square (LS) regression line* is the line $y = a + bx$ that minimizes the sum of squared errors:

$$\sum_i (y_i - \widehat{y_i})^2 = \sum_i (y_i - a - bx_i)^2$$

of which the slope and intercept are

$$b = \text{slope} = r \cdot \frac{\text{SD}_y}{\text{SD}_x}, \quad a = \text{intercept} = \bar{y} - \text{slope} \cdot \bar{x}$$

- Statistical interpretation of
  - the slope is the average change in y, associated with a unit increase in x.
  - the intercept is predicted value of y when $x = 0$ (only meaningful if 0 is a possible value of x)

Summary (Cont'd)

- Extrapolation is usually unreliable
- For LS regression, residuals add up to zero, and have 0 correlation with the explanatory variable x.
- $R^2 =$ R-squared $= r^2 =$ proportion of variation in the response y that can be explained by the explanatory variable
- Regression treats x and y differently.
  The LS regression line that predicts y from x is NOT the LS regression line that predicts x from y.
  The LS regression line that predicts y from x can only predict y from x, not x from y.