

MATH 1051H - Lecture 06

Linear Models

Kat Clark & Wesley Burr

Trent University

Today

- Lines of best fit
- Our first statistical model!

Readings

Chapter 7 of **Introduction to Modern Statistics (2e)**.

Our Coffee / Sleep Data

Let's load in the Zoom data:

```
1 library(readr)
2 coffee <- read_csv("data/CoffeevsSleep.csv", skip = 10)
```

Now the variable names are the whole question, so let's shorten those:

```
1 coffee <- coffee |>
2   rename(coffee = `How much coffee do you drink per day on average? (Answer
3           sleep = `How much sleep do you get on average? (Answer your best g
4   select(coffee, sleep)
```

And I do some final cleaning (details not important):

```
1 coffee <- coffee |>
2   mutate(coffee = ifelse(coffee == "3L", 3000, coffee),
3           coffee = parse_number(coffee),
4           sleep = parse_number(sleep),
5           sleep = ifelse(sleep < 10, sleep*60, sleep))
```

Data Summary

```
1 coffee |> summary()
```

coffee		sleep	
Min.	: 0.0	Min.	: 60.0
1st Qu.:	0.0	1st Qu.:	360.0
Median	: 218.0	Median	: 420.0
Mean	: 239.7	Mean	: 417.9
3rd Qu.:	355.0	3rd Qu.:	465.0
Max.	: 3000.0	Max.	: 612.0
		NA's	: 3

Everybody answered the coffee question, but 3 people left sleep blank. If we use the `summarize()` function, this could lead to issues:

```
1 coffee |> summarize(avg_sleep = mean(sleep),
2                     sd_sleep = sd(sleep))
```

```
# A tibble: 1 × 2
  avg_sleep sd_sleep
  <dbl>    <dbl>
1      NA        NA
```

The NA Solution

We have some options to address this. Either, we can use the argument `na.rm = TRUE` for the mean and sd, *remove NA*

```
1 coffee |> summarize(avg_sleep = mean(sleep, na.rm = TRUE),
2                     sd_sleep = sd(sleep, na.rm = TRUE))
```

```
# A tibble: 1 × 2
  avg_sleep sd_sleep
  <dbl>     <dbl>
1    418.       76.6
```

but not all functions have this argument, so we can also filter:

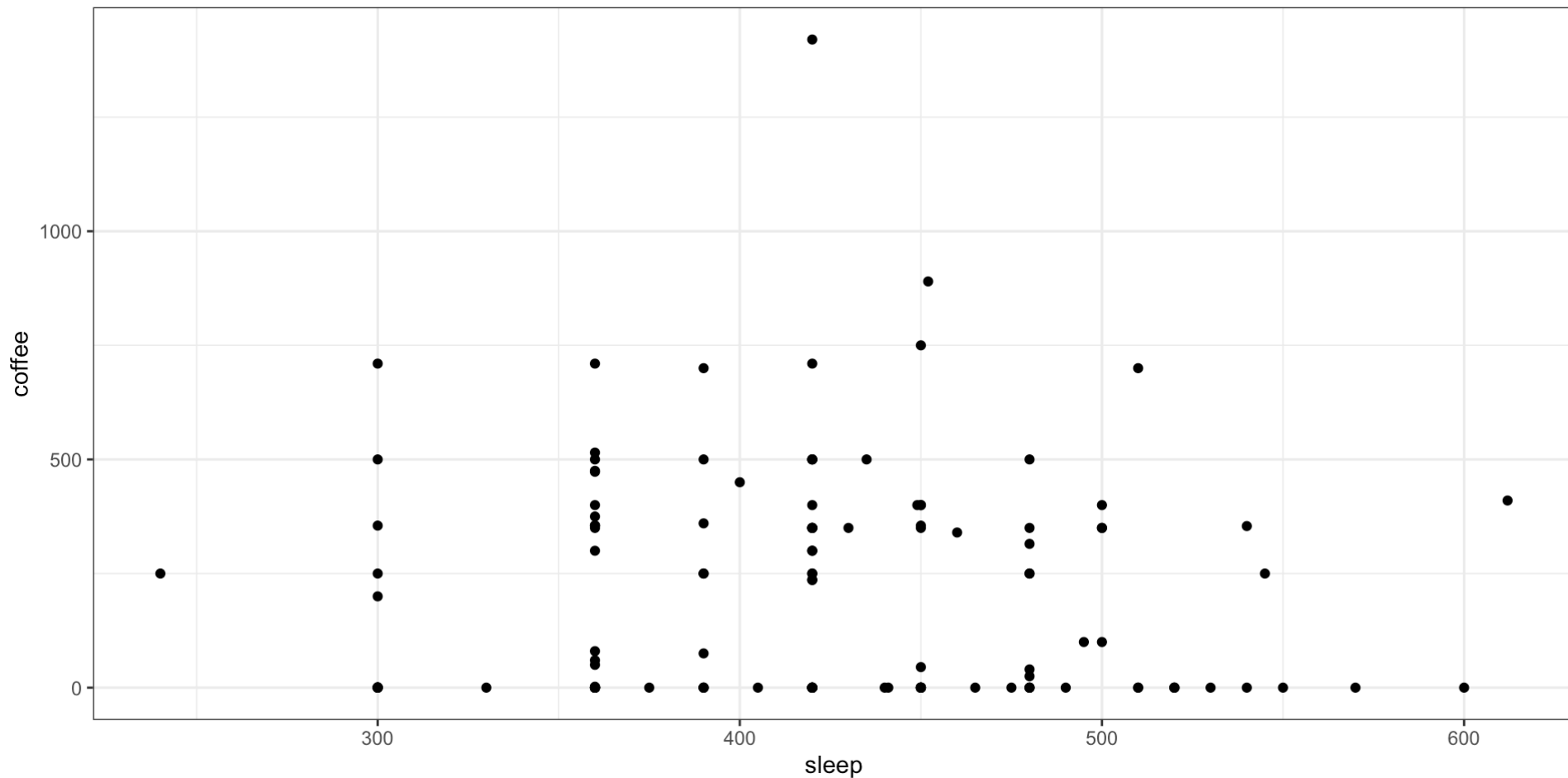
```
1 coffee |> ✓ filter(sleep > 0)
2   filter(!is.na(sleep)) |>
3   summarize(avg_sleep = mean(sleep, na.rm = TRUE),
4             sd_sleep = sd(sleep, na.rm = TRUE))
```

```
# A tibble: 1 × 2
  avg_sleep sd_sleep
  <dbl>     <dbl>
1    418.       76.6
```

Coffee / Sleep Plot

Now, let's plot! Our choice of response/explanatory is kind of arbitrary here.

```
1 coffee |> ggplot(aes(x = sleep, y = coffee)) + geom_point() + theme_bw()
```



Coffee/ Sleep Correlation

There is a **very weak** negative correlation here.

```
1 coffee |> summarize(correlation = cor(sleep, coffee))
```

```
# A tibble: 1 × 1  
  correlation  
    <dbl>  
1    -0.0684
```

- Unfortunately, this is not a great example of a linear relationship for the rest of our slides, so we'll have to go back to our poverty example.
- But, let's try again before Monday's class

Let's Do More Science! (Homework)



- We are going to investigate whether there is a **linear relationship** between the *number of chapters* in a book and the *number of pages* in the book.
- Please choose any book you have access to. It can be a textbook, novel, nonfiction book, cookbook, graphic novel, etc.

Deadline: Friday at Noon

- You may submit data for as many different books as you like, but each submission should represent a different book.
- For each book, enter: the number of chapters; the number of pages; the name of book.
- This activity is *not* required, but it's a fun way to make some data to use on Monday.

Fitting a line by least squares regression

The Idea

- We want to fit a straight line of the form:

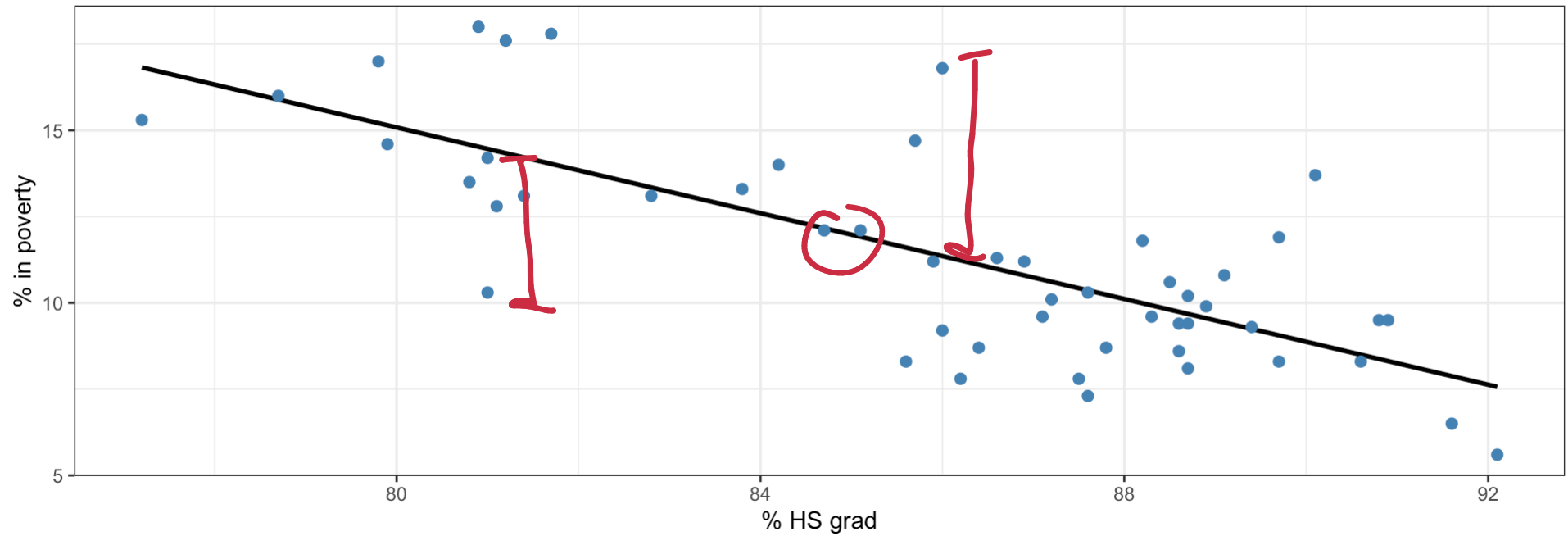
$$\hat{y} = b_0 + b_1x$$

$$y = mx + b$$

- This line should be as “close” to each point as possible.
 - How do we define “close” ?
- The **model** we will use is:

$$Y_i = \beta_0 + \beta_1x_i + \epsilon_i$$

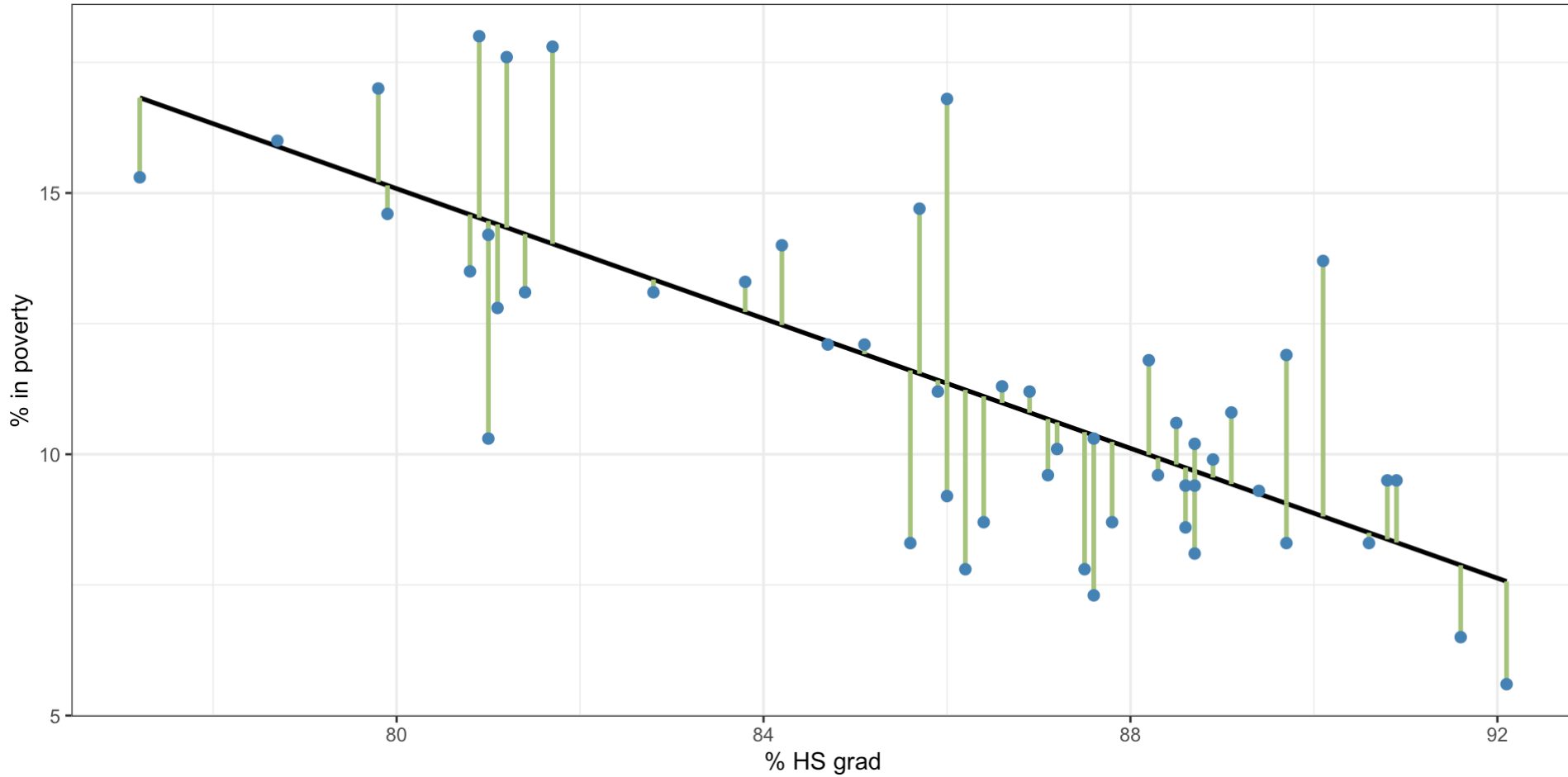
The Data



- We have data from 50 states (plus DC) on poverty and high school graduation rates.
- The black line is the “best” line that describes the relationship.
- How do we calculate this line?

Residuals

Residuals are the leftovers from the model fit: $\text{Data} = \text{Fit} + \text{Residual}$



Residuals (cont.)

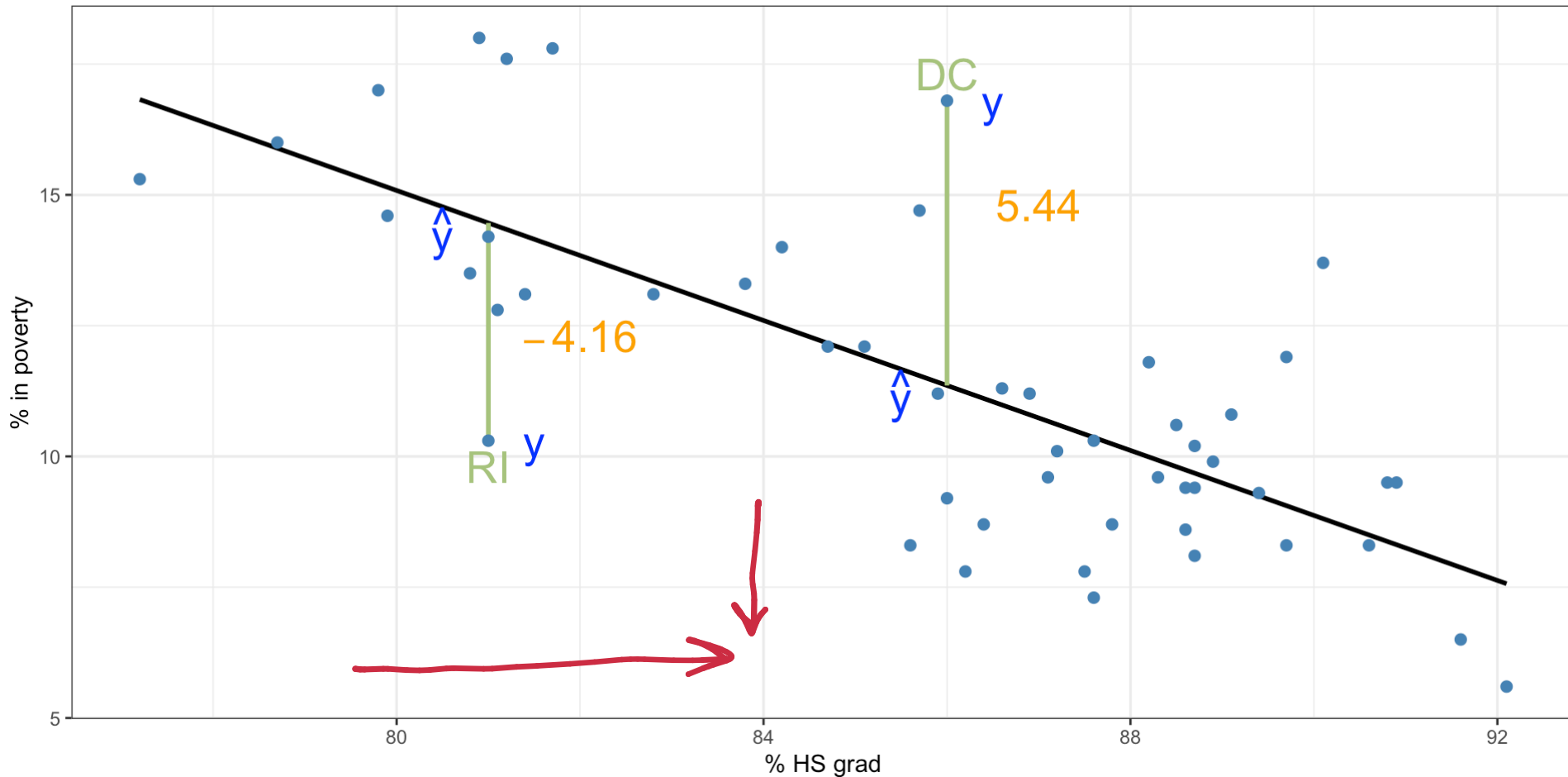
Formally, residuals are the difference between the observed values of the response (y_i) and predicted values of the response from the model ($\hat{y}_i$).

$$e_i = y_i - \hat{y}_i$$

↑
true, observed value

← predicted based on line

Specific Residuals



- % 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

What is “close”?

- We want a line that has small residuals:
 - 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 – known as **least squares**

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

- Option 3 ... 100: actual topics of research!
- Why least squares?
 - Under the assumption that $\epsilon_i \stackrel{\text{iid}}{\sim} N(\mu, \sigma_\epsilon)$, the least-squares estimates are maximum likelihood!
 - Most commonly used
 - Easier to compute by hand and using software
 - In many applications, a residual twice as large as another is usually more than twice as bad

The least squares line

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

Handwritten annotations in red:

- An arrow points from the word "estimate" to $\hat{y}$.
- An arrow points from the word "intercept" to β_0 .
- An arrow points from the word "slope" to $\beta_1 x$, which is circled.

Notation:

- Intercept:
 - Parameter: β_0
 - Point estimate: b_0
- Slope:
 - Parameter: β_1
 - Point estimate: b_1

The least squares line

```
1 mod <- lm(Poverty ~ Graduates, data = poverty)
2 summary(mod)
```

$y \sim x$

Call:

```
lm(formula = Poverty ~ Graduates, data = poverty)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.1624	-1.2593	-0.2184	0.9611	5.4437

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	64.78097	6.80260	9.523	9.94e-13 ***
Graduates	-0.62122	0.07902	-7.862	3.11e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.082 on 49 degrees of freedom

Multiple R-squared: 0.5578, Adjusted R-squared: 0.5488

F-statistic: 61.81 on 1 and 49 DF, p-value: 3.109e-10

Slope

The **slope** of the regression can be calculated as

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

but we will just use R , and identify it from the summary table:

```
1 mod |> coefficients()
```

(Intercept)	Graduates
64.7809658	-0.6212167

(so $b_1 = -0.62$)

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

Intercept

The **intercept** is where the regression line intersects the y -axis. The calculation of the intercept uses the fact the regression line always passes through $(\bar{x}, \bar{y})$:

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

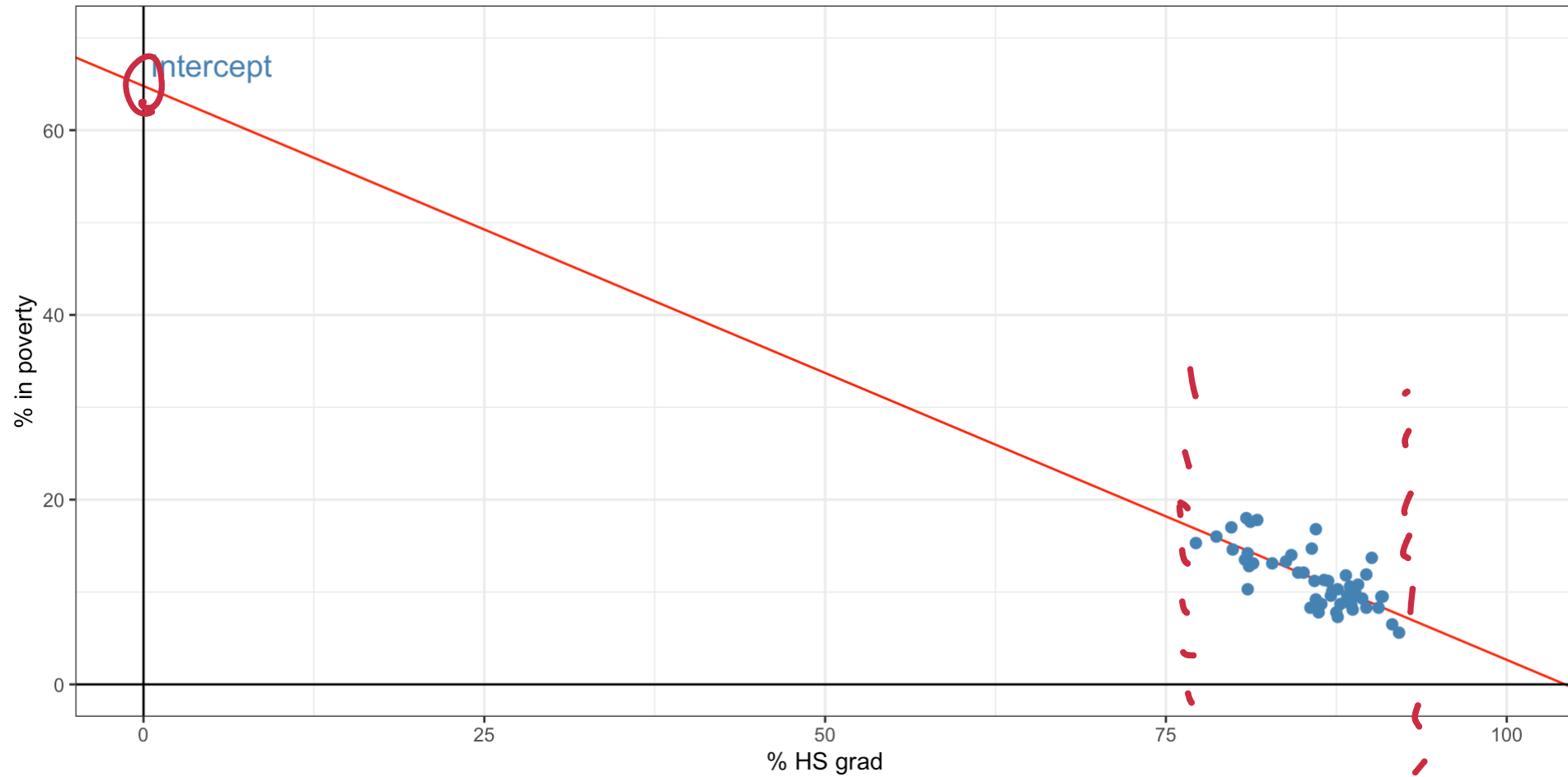
but again, we will just use R and identify it:

```
1 mod |> coefficients()
```

```
(Intercept)    Graduates  
64.7809658    -0.6212167
```

(so $b_0 = 64.78$).

Intercept



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

- For each % point ~~increase~~ in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.
- ~~•~~ For each % point decrease in HS graduate rate, % living in poverty is expected to increase on average by 64.68%.
- Having no HS graduates leads to 64.68% of residents living below the poverty line.
- States with no HS graduates are expected on average to have 64.68% of residents living below the poverty line.
- ~~•~~ In states with no HS graduates, % living in poverty is expected to increase on average by 64.68%.

More on the intercept

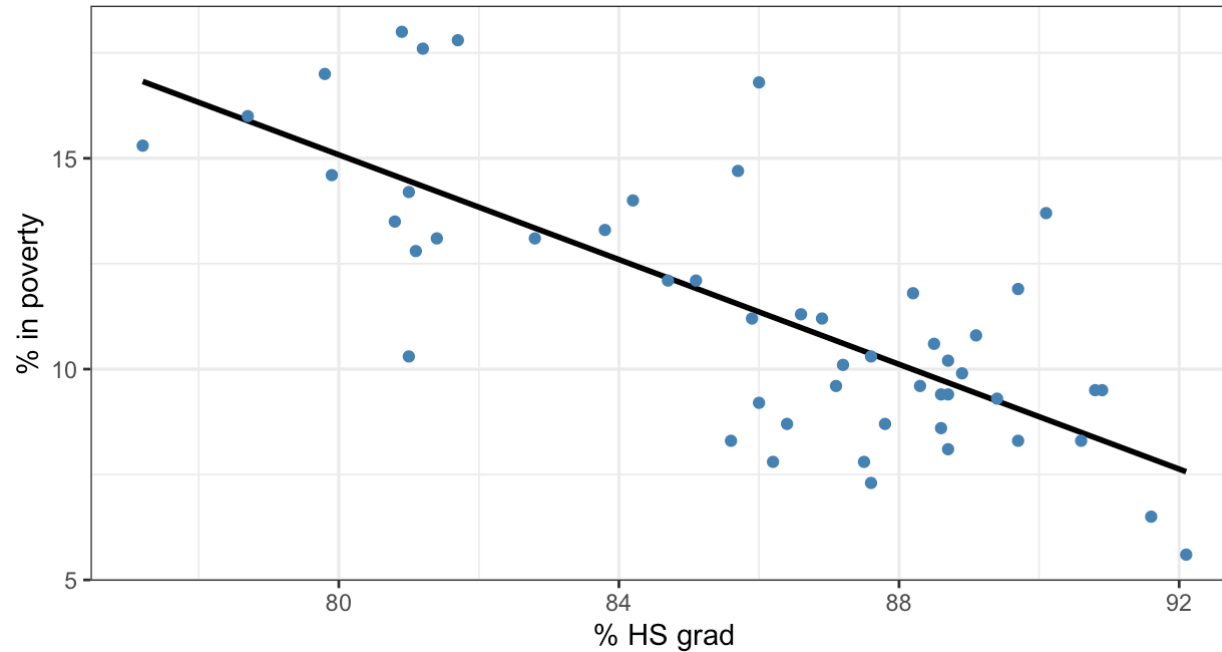
Since there are no states in the dataset with no HS graduates, the intercept is of no interest, not very useful, and also not reliable since the predicted value of the intercept is so far from the bulk of the data.

(see previous figure)

But we need it to define the line,

Regression line

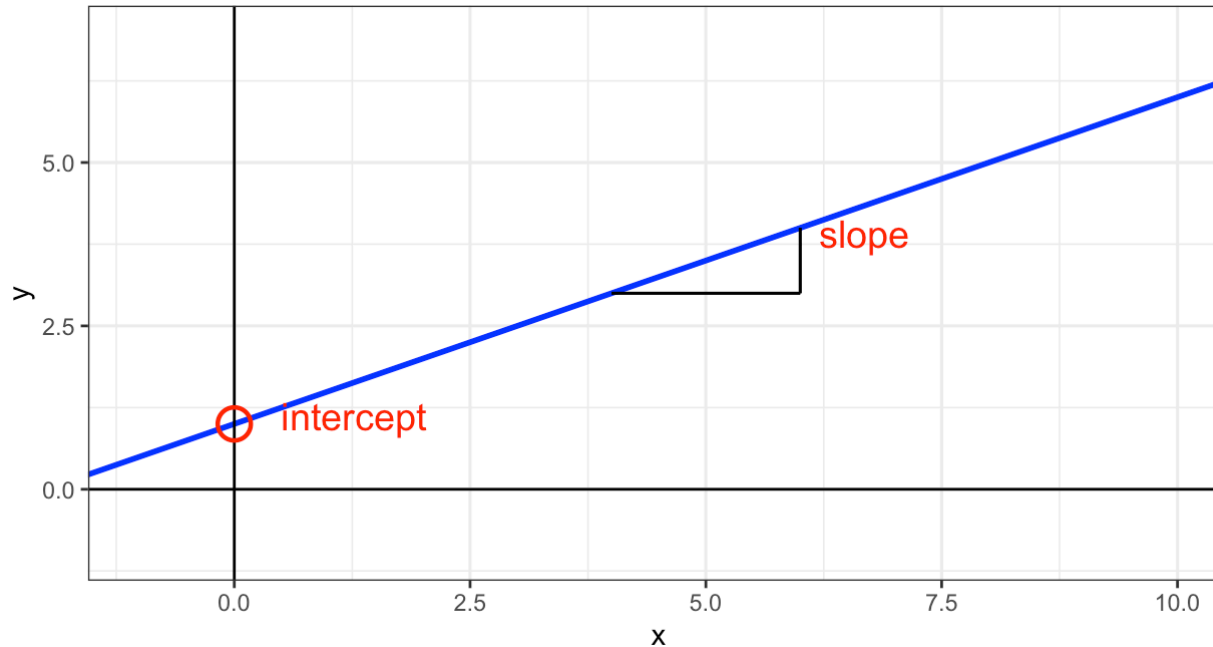
```
1 ggplot(poverty, aes(x = Graduates, y = Poverty)) +
2   geom_smooth(method = "lm", se = FALSE, col = "black") +
3   geom_point(col = "steelblue") + theme_bw() +
4   labs(y = "% in poverty", x = "% HS grad")
```



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

Interpretation of slope and intercept

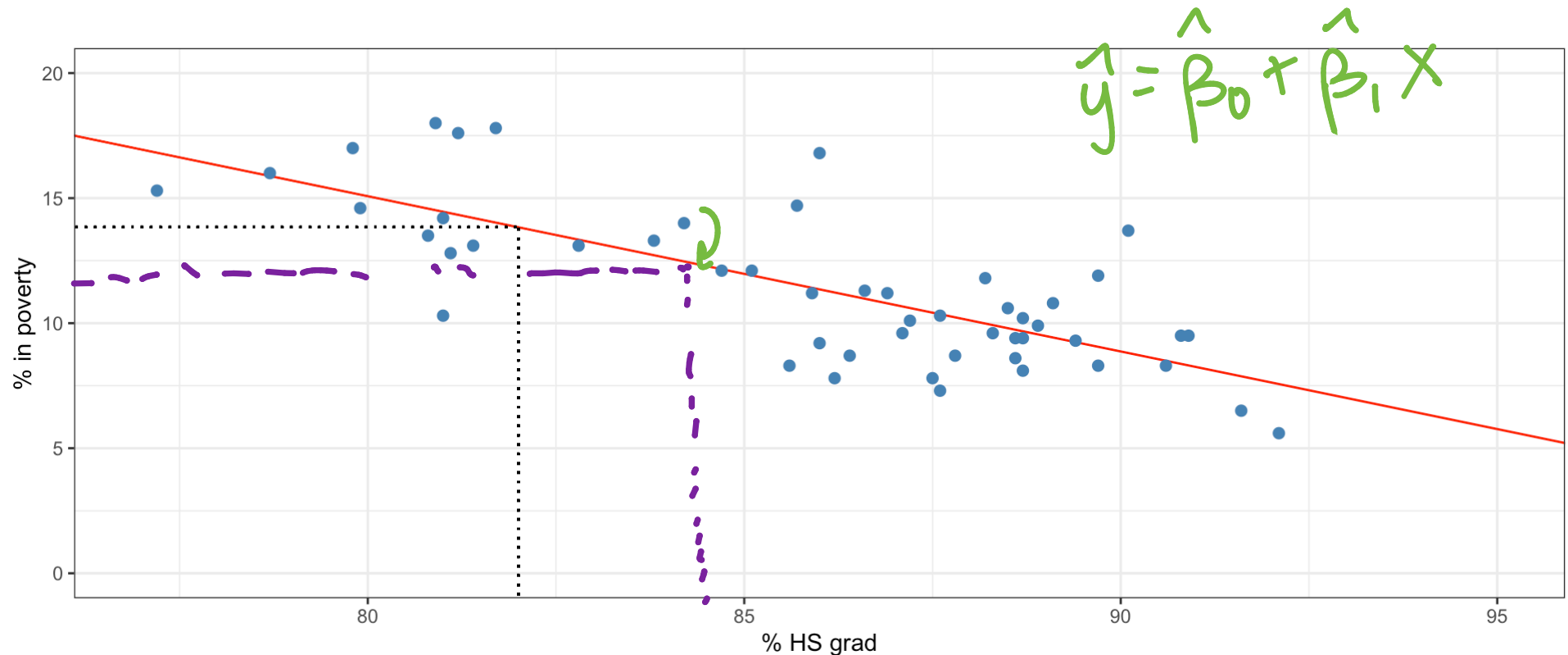
- **Intercept:** When $x = 0$, y is expected to equal the intercept.
- **Slope:** For each unit in x , y is expected to increase / decrease on average by the slope.
↳ increase in one unit



Prediction & extrapolation

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.



What is predicted % poverty for a state with a 84% graduation rate?

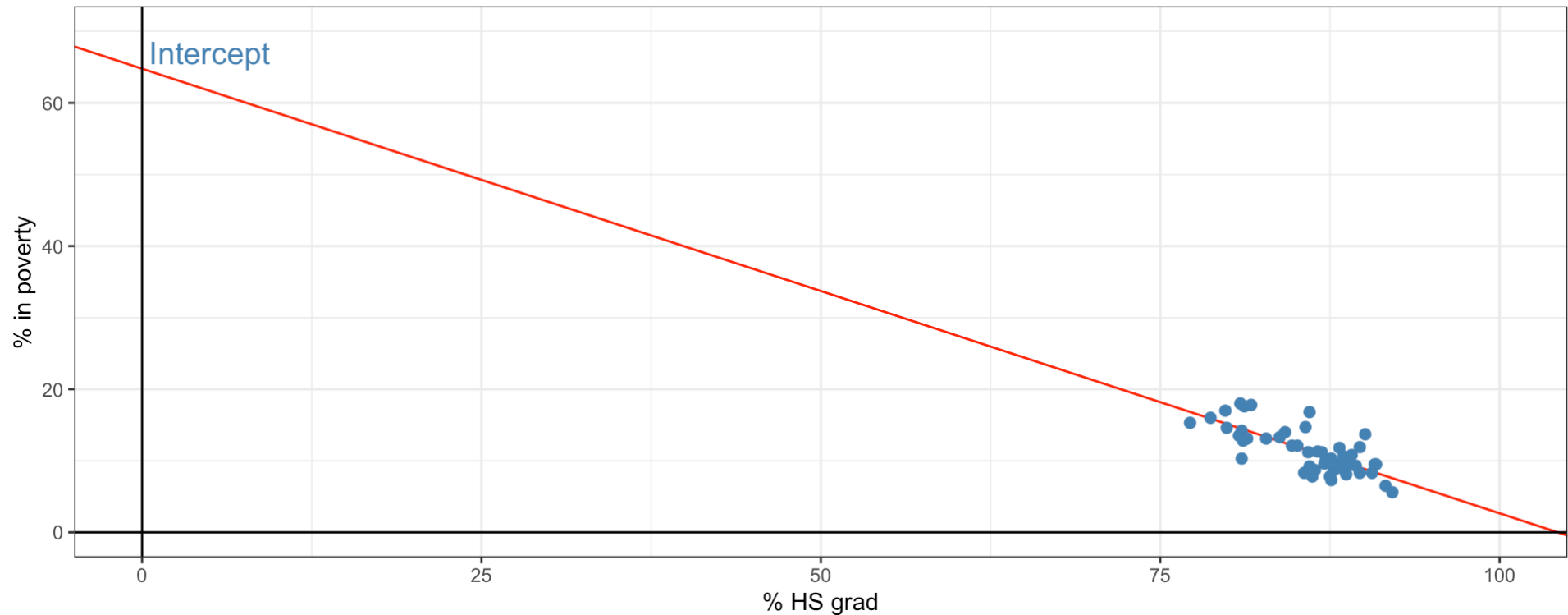
$$\hat{y} = 64.68 - 0.62 \times (84) \\ = 12.6\%$$

What is predicted % HS graduation for poverty = 15%?

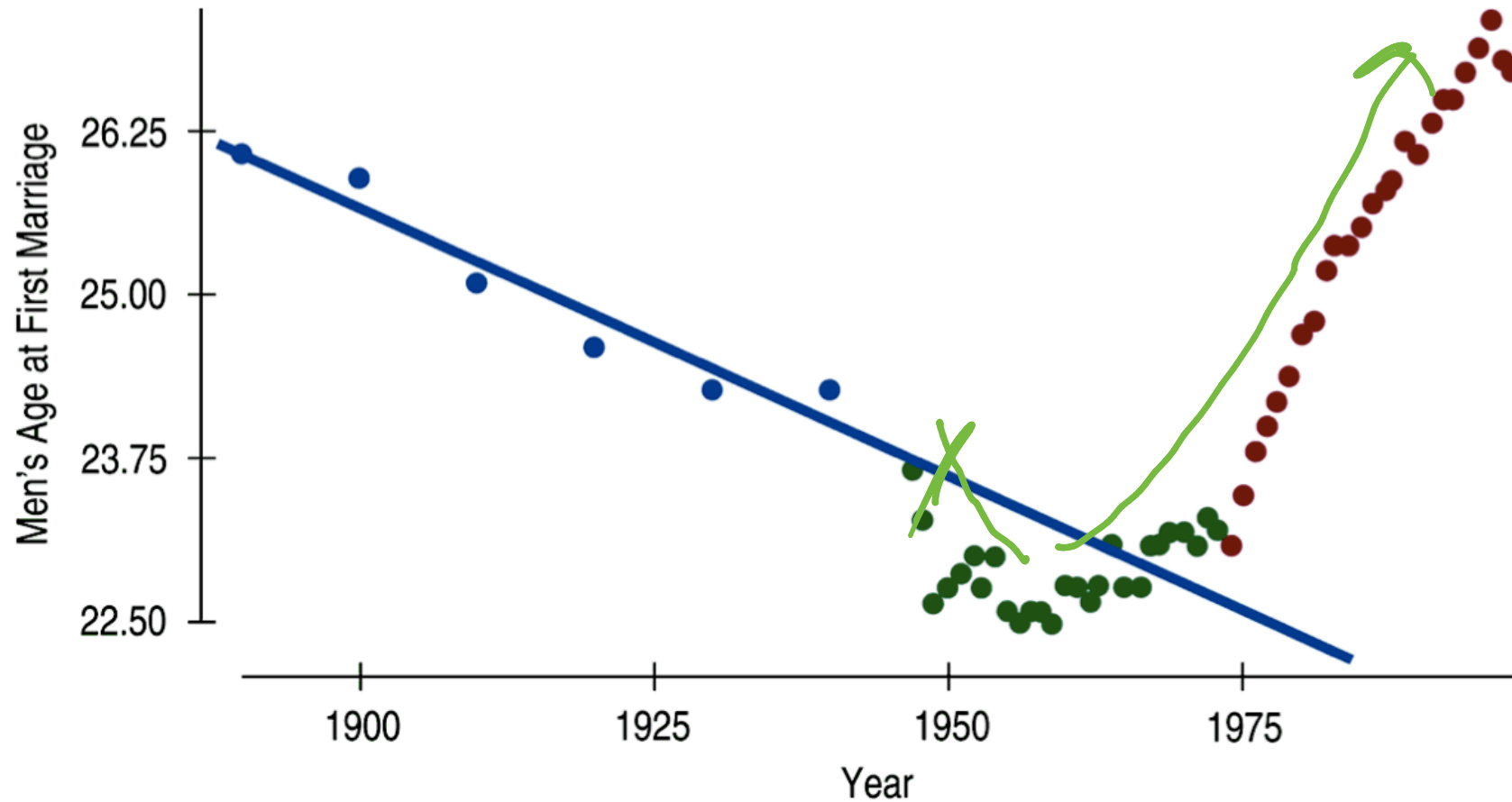
$$15 = 64.68 - 0.62 \times \quad \rightarrow \quad \begin{array}{r} 15 - 64.68 \\ -0.62 \\ \hline = 80.13\% \end{array}$$

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



Examples of extrapolation

Momentous sprint at the 2156 Olympics?

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

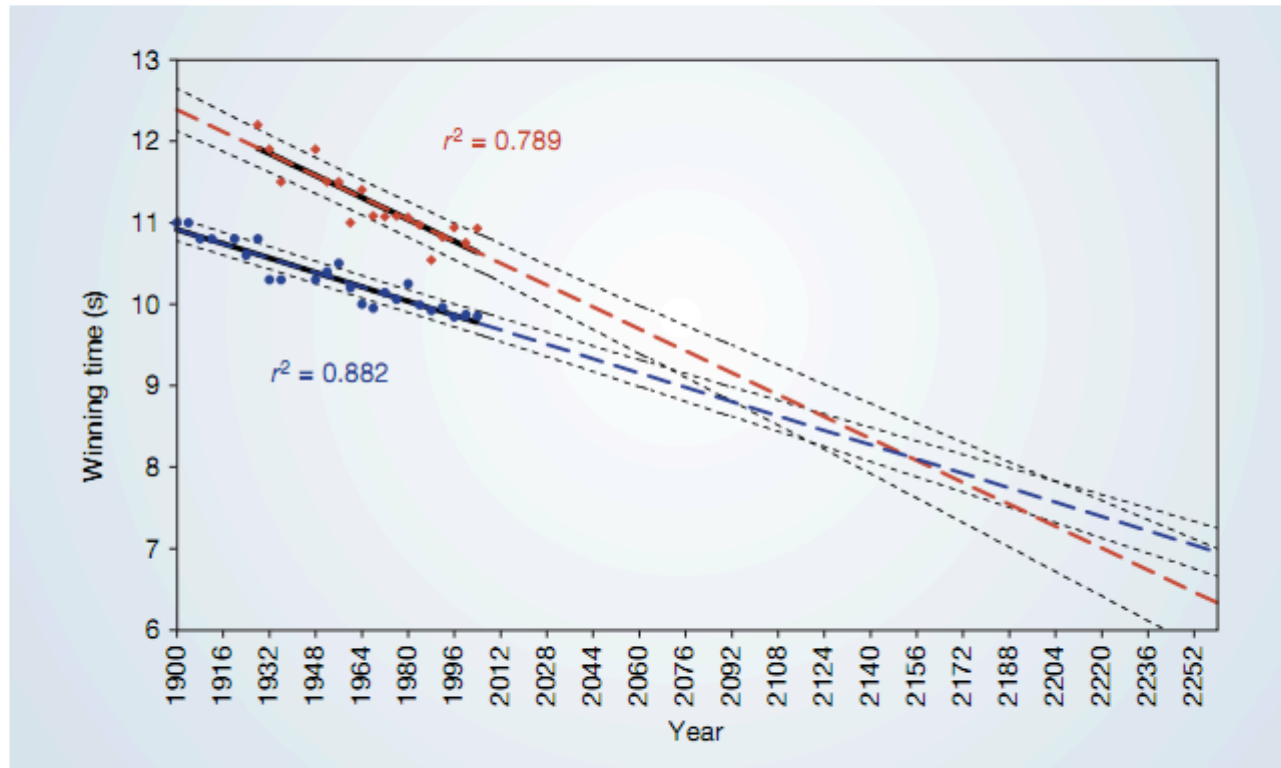


Figure 1 The winning Olympic 100-metre sprint times for men (blue points) and women (red points), with superimposed best-fit linear regression lines (solid black lines) and coefficients of determination. The regression lines are extrapolated (broken blue and red lines for men and women, respectively)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



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

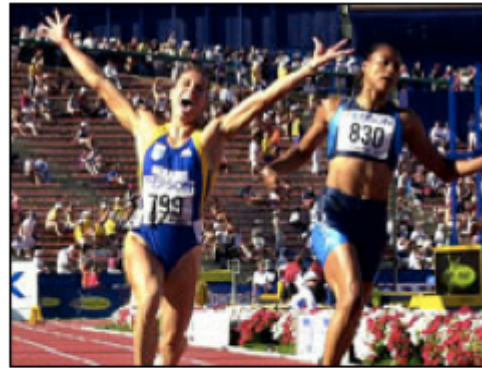
Last Updated: Thursday, 30 September, 2004, 04:04 GMT 05:04 UK

 [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.

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



Women are set to become the dominant sprinters

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

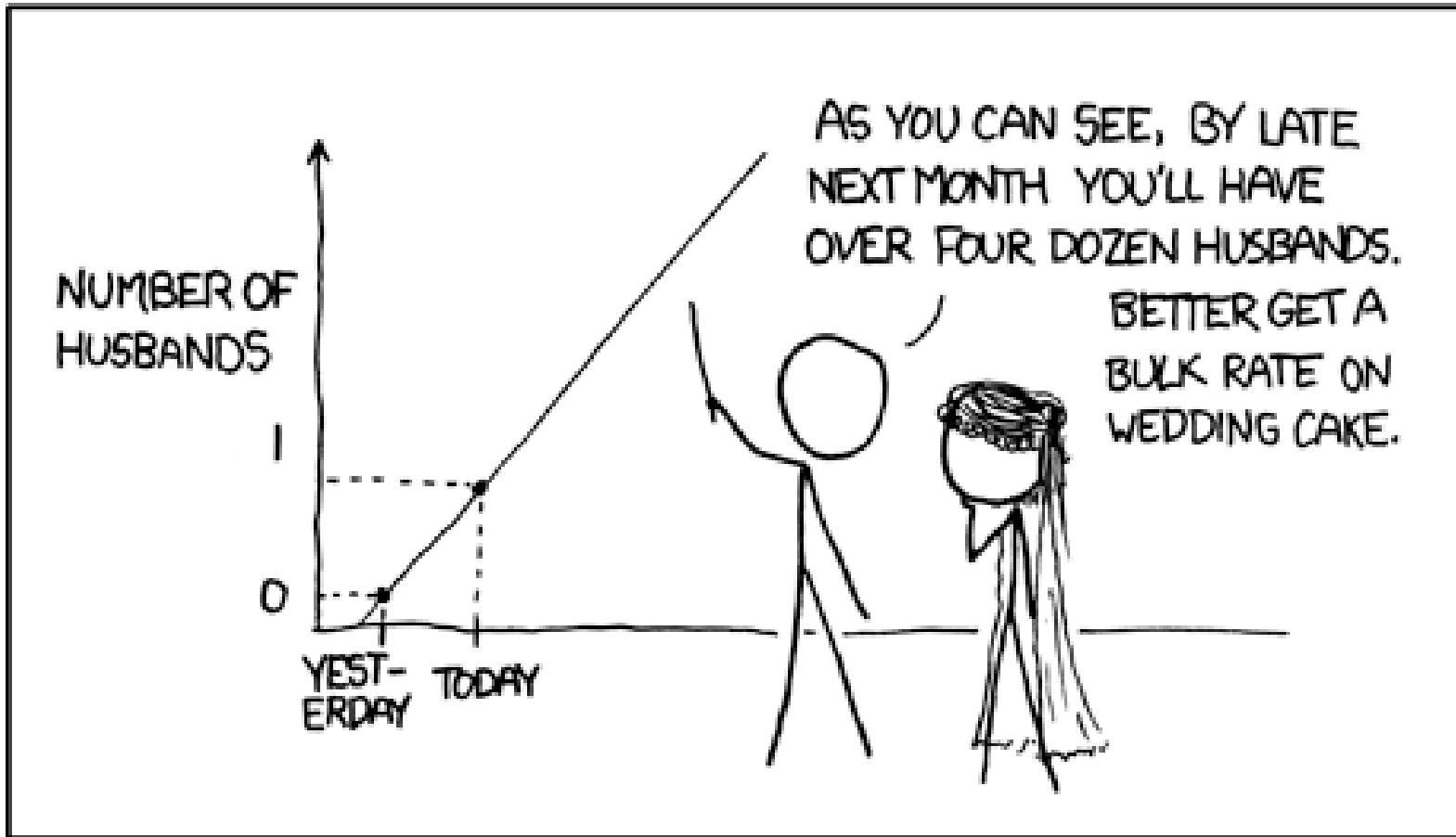
The study, comparing winning times for the Olympic 100m since 1900, is published in the journal Nature.

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 extrapolation¹

MY HOBBY: EXTRAPOLATING



To Come

We now have our first statistical model. In the coming lectures, we will talk about what this model means, how we use it, and how we can determine uncertainty in the model (since it is statistical, after all!).