Simple Linear Regression

Grinnell College

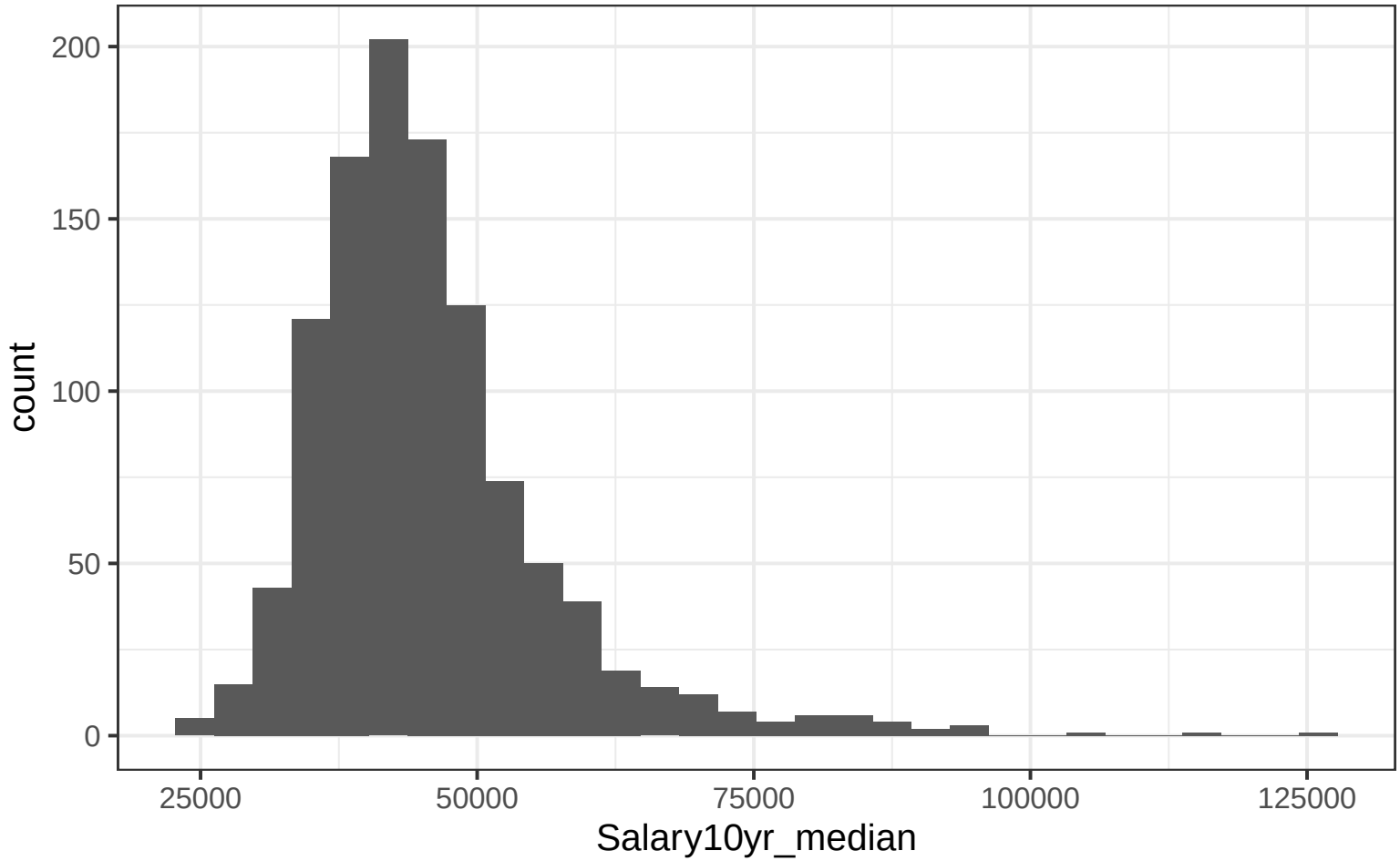
Feburary, 2026

Review

- Scatterplot descriptions
  - form, strength, direction, outliers
- Pearson's correlation (r)
  - strength and direction of linear relationship for 2 quant. variables
- Spearman's correlation (ρ)
  - strength and direction of *monotone* relationship
  - more robust to outliers

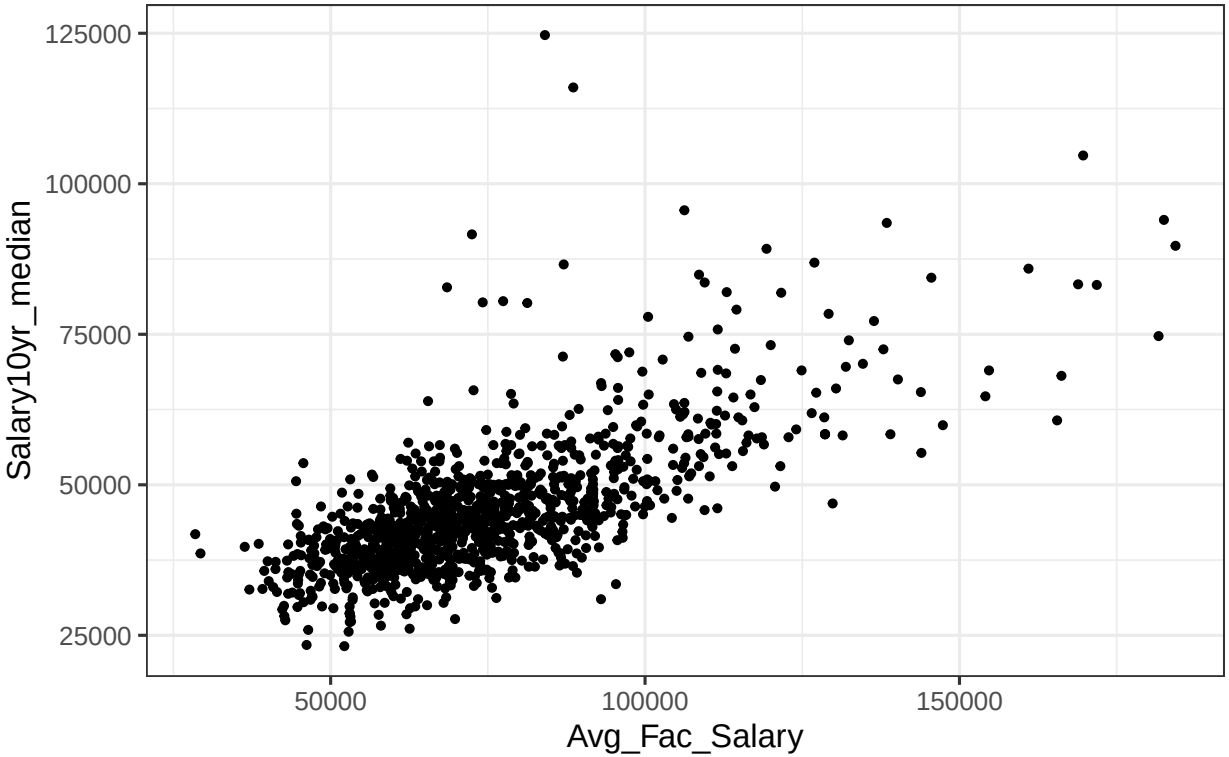
Motivation

If I asked you to guess your income after ten years, how would you guess?



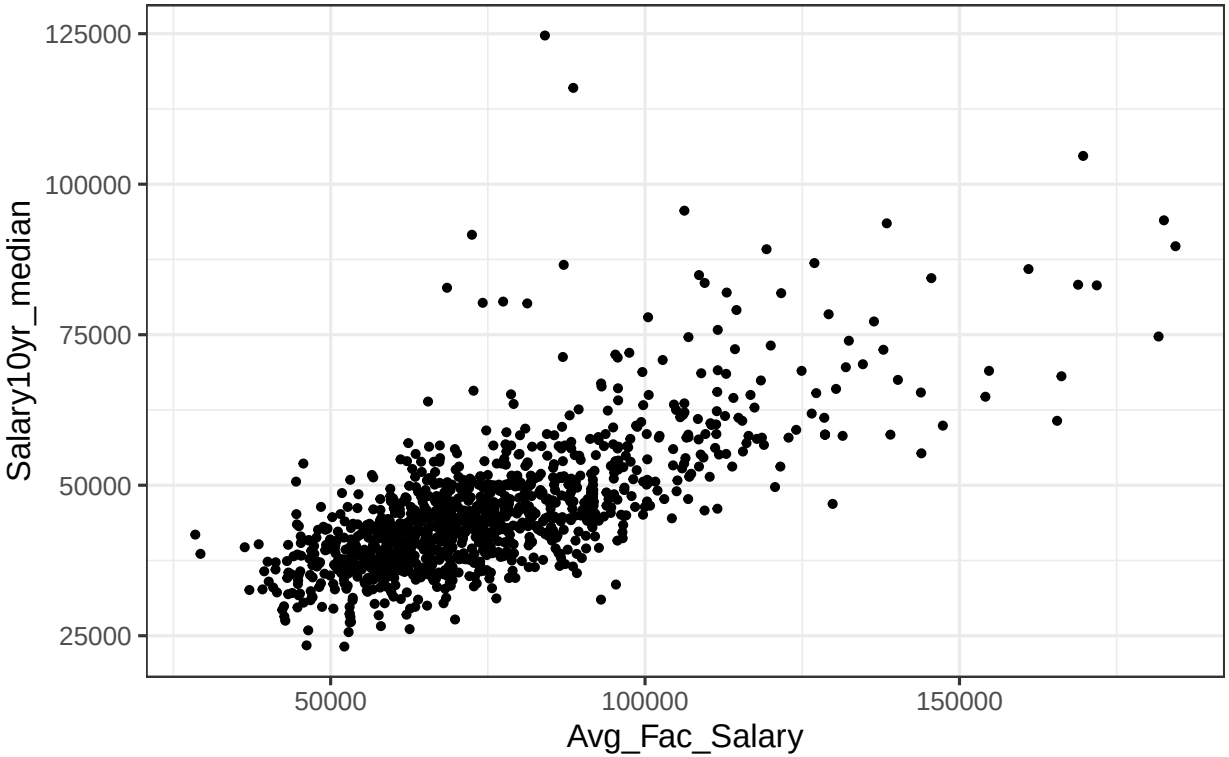
Motivation

If I told you my salary, how would you guess your (future) income?



Motivation

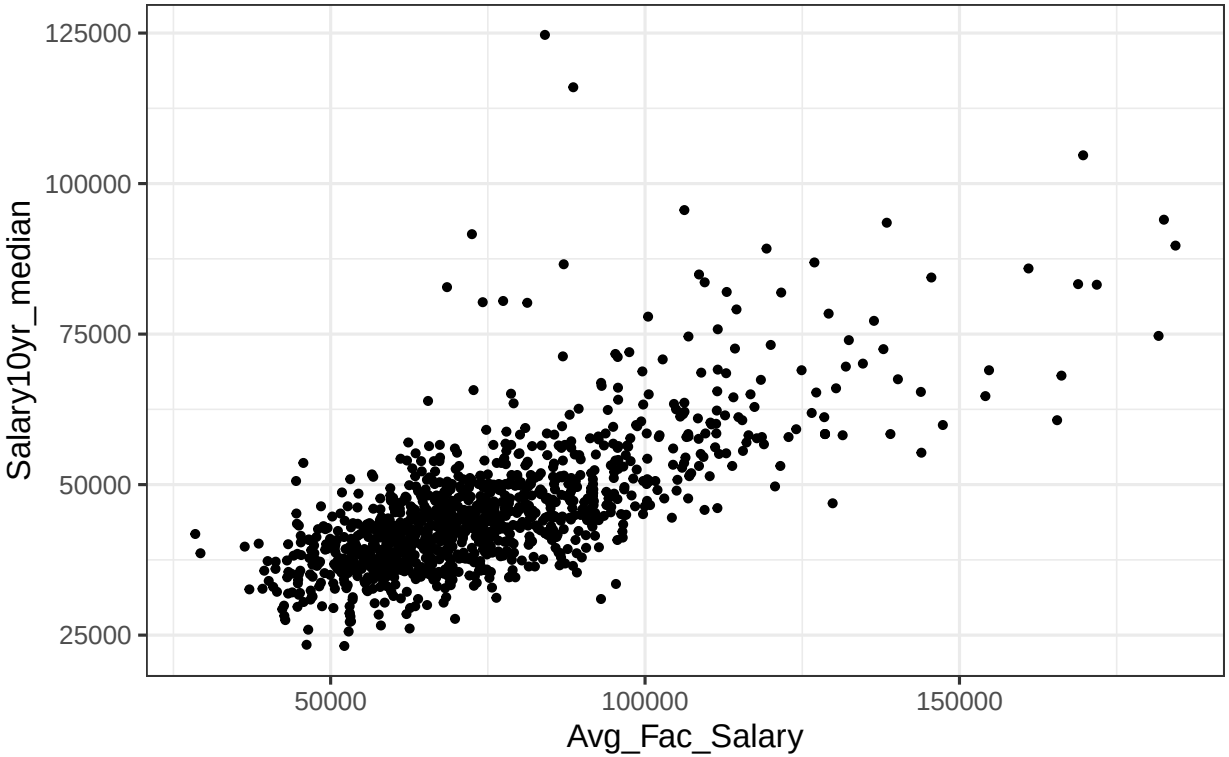
If I told you my salary, how would you guess your (future) income?



Linear Regression allows us to do this formally

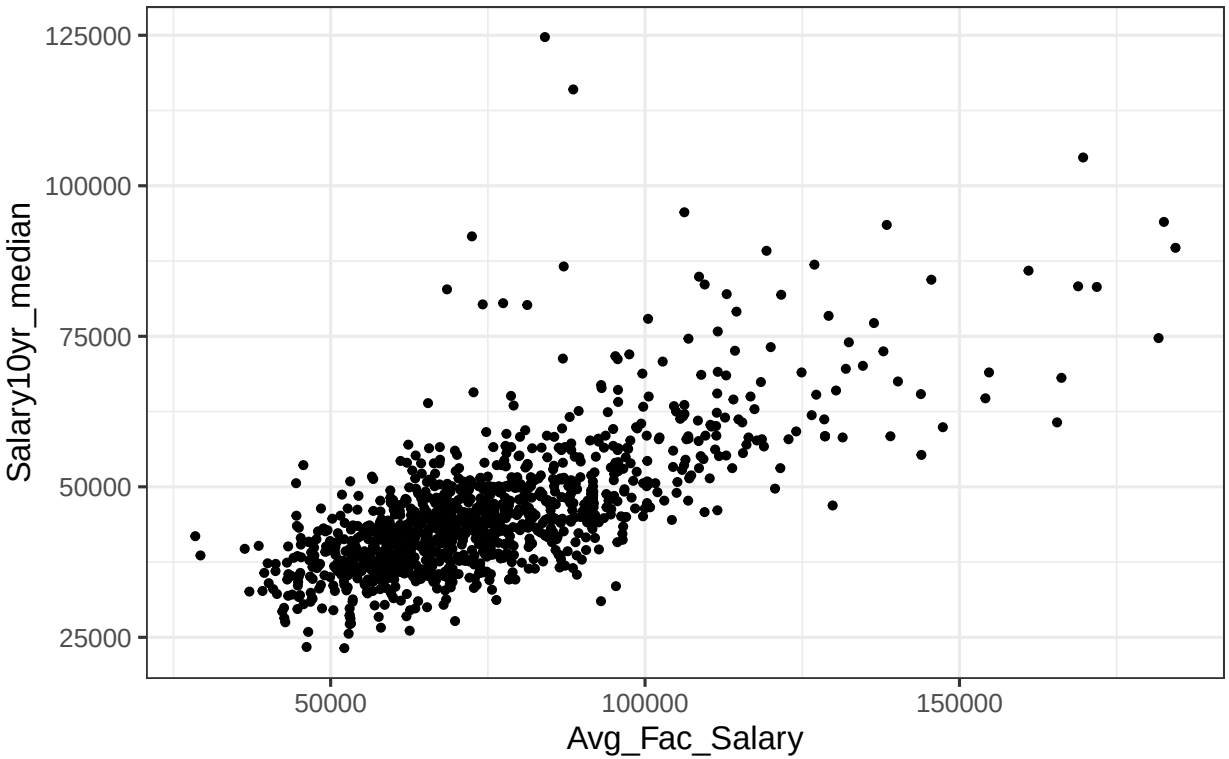
Correlation, Causation Review

Should you all tell the administration to raise my salary so your future income increases?



Correlation, Causation Review

Should you all tell the administration to raise my salary so your future income increases?



Yes!! But it won't actually increase your income. They are correlated but one doesn't cause the other (at least directly).

Basic Idea

Regression is how we model data; for us it's the "best fit line"

Two Main Goals:

- Use the regression/our best fit line(s) to describe the relationship between the explanatory variable(s) and the response variable
  - Science!
  - Hypothesis testing
- Use the explanatory variable(s) to predict the response variable
  - Machine Learning/AI stuff
  - Business/finance investments
  - Planning around weather

Notation

- The variable being predicted is the *response* (aka "variable of interest")
  - Usually denoted as y
- the variable we are using to do the prediction/explanation is the *explanatory variable* (aka "covariate" or occasionally "predictor")
  - Usually denoted as x or X
- The estimates themselves are usually denoted with a "hat"
  - $\hat{y}$ is our predicted response
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ are our estimated intercept and slope of the regression line (more in a second)

Notation Comparison

Statisticians use different symbols to write out a line than what you probably saw in HS algebra

Algebra

$y = mx + b$

m = slope: change in y over the change in x (rise / run)

b = intercept: value where the line cross the y-axis

All points fall exactly on the line

Statistics

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

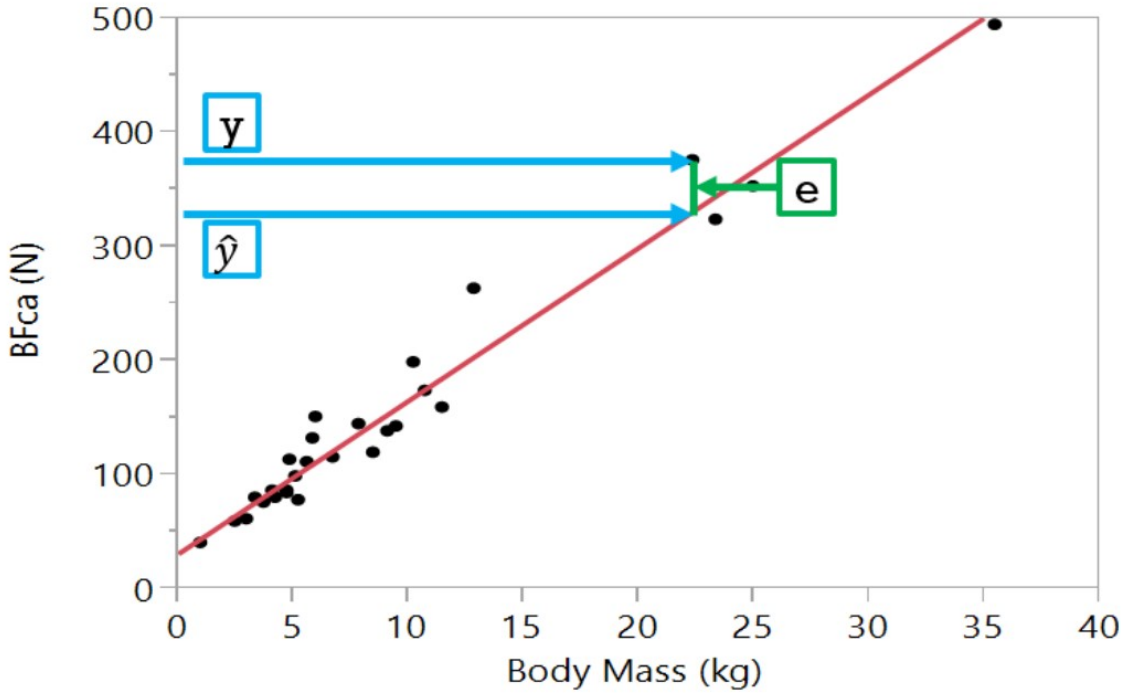
β_1 = slope

β_0 = intercept

Not all of our data points will exactly on the line $\rightarrow$ variability

How it works

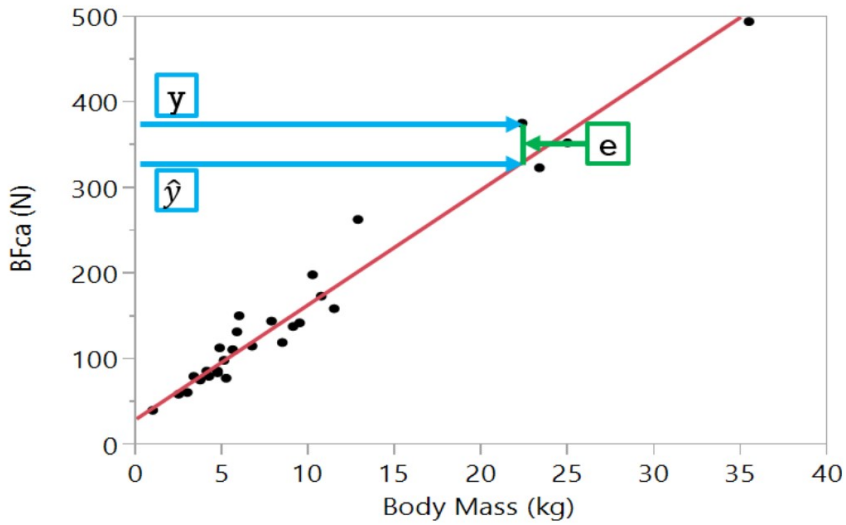
A regression line for the canidae data set predicting bite force (response) using body mass (explanatory)



- y's denote the values of the datapoints for the response variable
- points on the line are predicted values for the y's, denoted as $\hat{y}$
  - $\hat{y}$ are ALWAYS on our best-fit-line
- residual: difference between data and predictions (e = y - $\hat{y}$)

How it works

The **regression line** is the line that best fits through the data



- Need to define "best"
- Optimality critera: minimizes sum of squared residuals $\sum e_i^2$
- *Least Squares Regression* is another, more explicit name for this

Some Formulas

- $\hat{y} = \beta_0 + \beta_1 X$ (**regression equation**)
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$ (**estimated regression equation**)
- $\hat{\beta}_1 = (\frac{s_x}{s_y})r$ (estimated slope)
- $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$ (estimated intercept)
- e = y - $\hat{y}$ (**residual**)

Regression Line vs Estimated Regression Line

What is the difference between these two? Why do we have two?

- $\hat{y} = \beta_0 + \beta_1 X$ (**regression equation**)
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$ (**estimated regression equation**)

Regression Line vs Estimated Regression Line

What is the difference between these two? Why do we have two?

- $\hat{y} = \beta_0 + \beta_1 X$ (**regression equation**)
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$ (**estimated regression equation**)

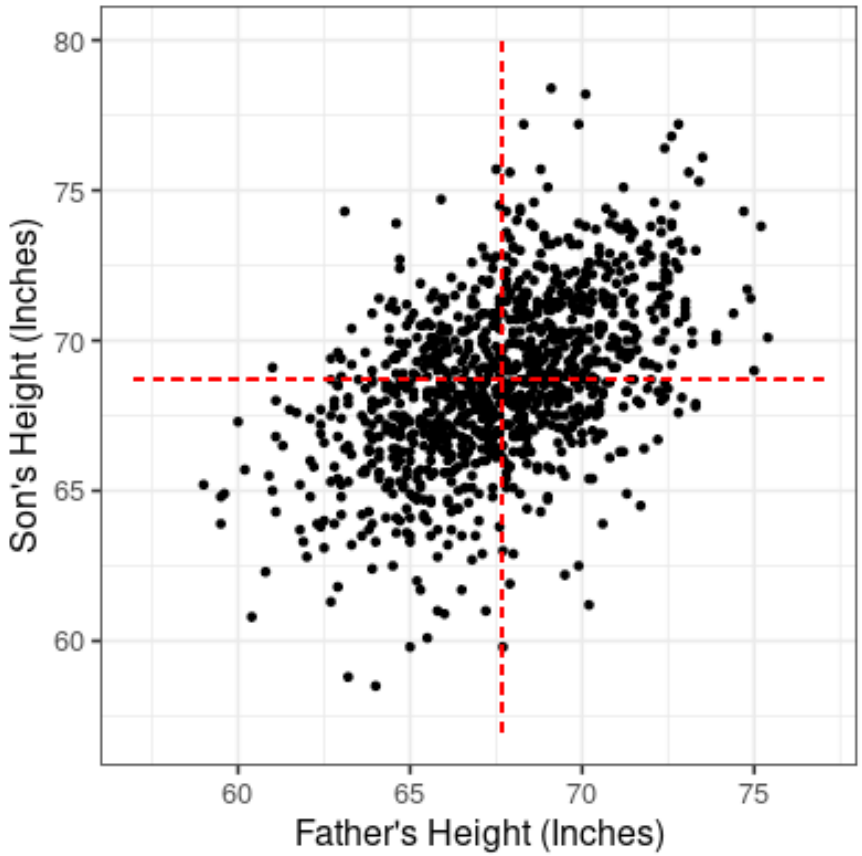
β_0 and β_1 are population parameters which means we almost never know them. Instead we have to estimate them using our sample.

Again, ˆ (called hat) means estimated

Pearson's Height Data

	Mean	Std.Dev.	Correlation (r_{xy})
Father	67.68	2.74	0.501
Son	68.68	2.81	

Father	Son
65.0	59.8
63.3	63.2
65.0	63.3
65.8	62.8
61.1	64.3
63.0	64.2
⋮	⋮



Pearson's Height Data

We could calculate our regression line using info from this table.

	Mean	Std.Dev.	Correlation (r_{xy})
Father	67.68	2.74	0.501
Son	68.68	2.81	

Regression equation:
$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$

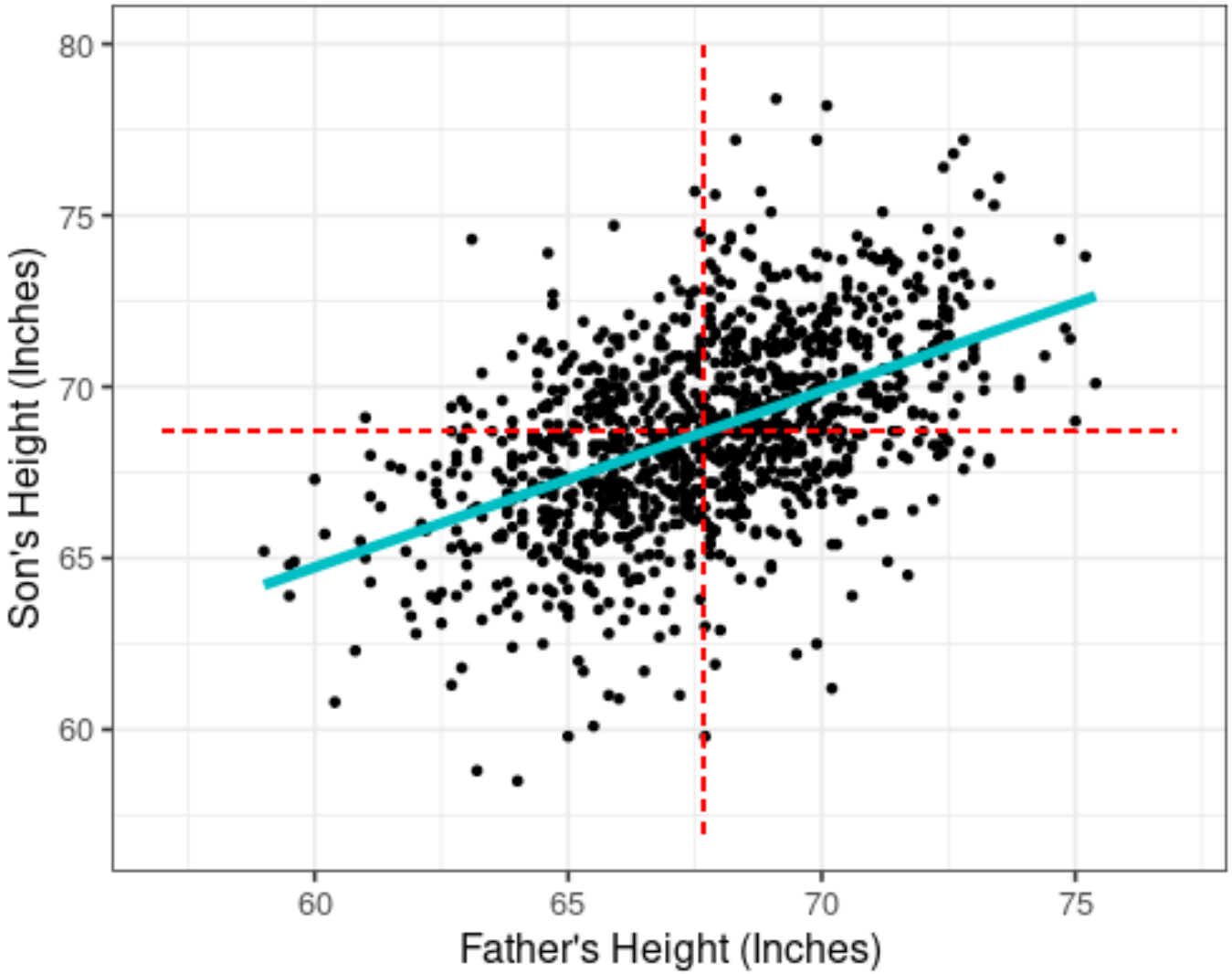
$$\hat{\beta}_1 = (\frac{s_x}{s_y})r$$
$$= (\frac{2.81}{2.74})0.501 = 0.514$$

```
> heights <- read.csv("Pearson.tsv", sep = "\t")
> fit <- lm(Son ~ Father, heights)
> fit

Call:
lm(formula = Son ~ Father, data = heights)

Coefficients:
(Intercept)       Father
     33.893        0.514
```

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$
$$= 68.68 - 0.514 * 67.68 = 33.893$$

Pearson's Height Data – Plot Line

We can make R graph the line on our scatterplot.



Pearson's Height Data – Prediction

The formula for the regression line

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

can be expressed in context of our original data and our estimated values

$$\widehat{\text{Son's Height}} = 33.9 + 0.51 \times \text{Father's Height}$$

Given the Father's height, we can predict the son's height using this equation by plugging in a value for the father's height

Example: Predict the height of the son for a father with a height of 65in.

$$\widehat{\text{Son's Height}} = 33.9 + 0.51 \times 65.0 = ?$$

Pearson's Height Data – Prediction

The formula for the regression line

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

can be expressed in context of our original data and our estimated values

$$\widehat{\text{Son's Height}} = 33.9 + 0.51 \times \text{Father's Height}$$

Given the Father's height, we can predict the son's height using this equation by plugging in a value for the father's height

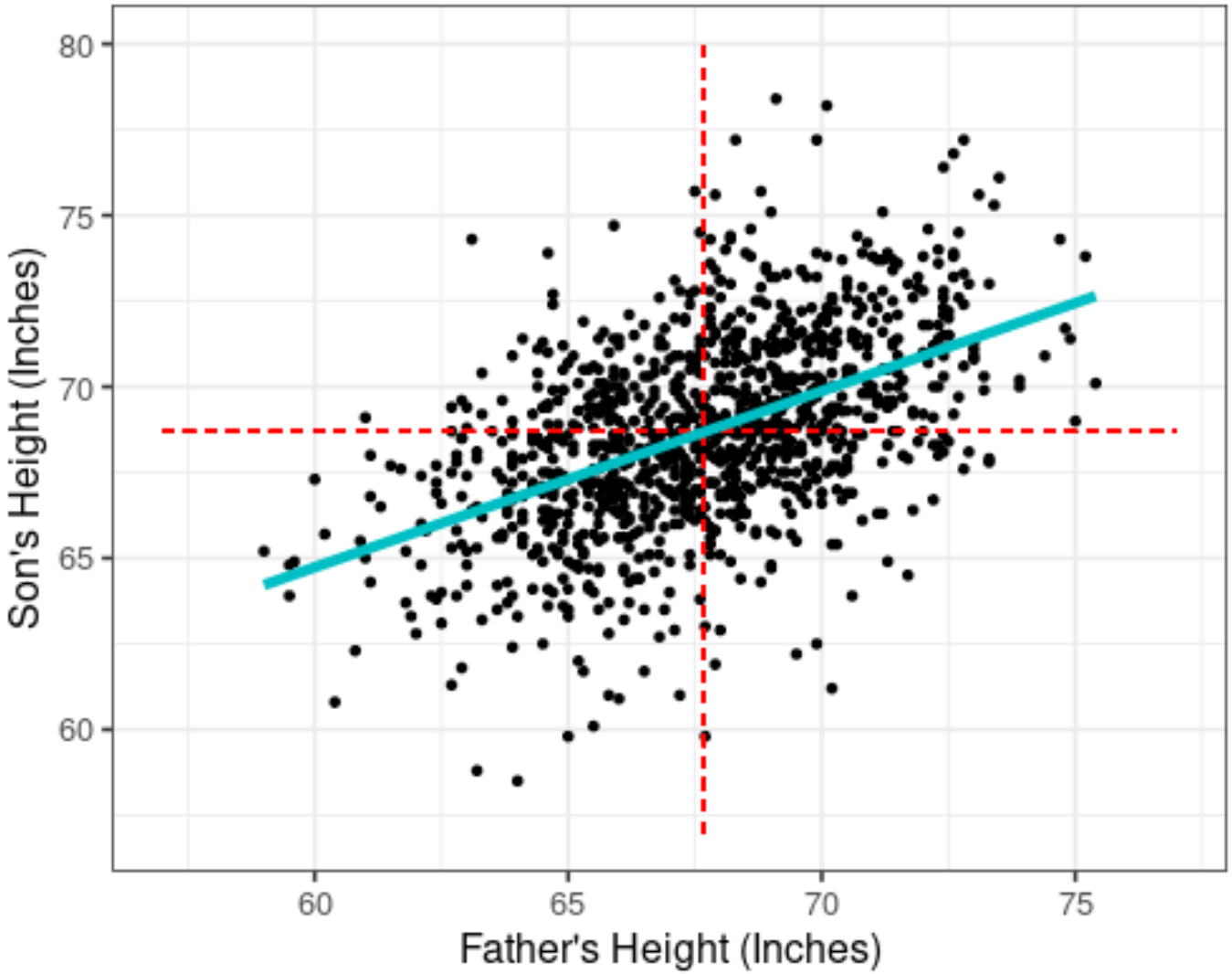
Example: Predict the height of the son for a father with a height of 65in.

$$\widehat{\text{Son's Height}} = 33.9 + 0.51 \times 65.0 = 67.30in.$$

Pearson's Height Data – Prediction

Predicted Son's Height = 67.30 inches for a father with height = 65in

- Check to see if our prediction makes sense on the graph



Residual

A **Residual** is the difference between an observed value and a prediction

- often labeled as **e** (e for "error", occasionally ϵ)
- e = y - $\hat{y}$

Interpretation: the residual tells us whether we have over- or under-predicted the values for the response variable in our data (and by how much)

- positive value $\rightarrow$ under-predicted
- negative value $\rightarrow$ over-predicted
- hard truth $\rightarrow$ I always forget which is which

Pearson's Height Data – Residual

In our data set, the first father had a height of 65 inches. We can calculate the residual for this father. We predicted the son's height to be 67.30 inches.

$$
\begin{aligned}
e &= y - \hat{y} \\
&= \text{observed value - predicted value} \\
&= 59.8in. - 67.30in. = -7.5in.
\end{aligned}
$$

Interpretation: We overpredicted the height of this particular son by 7.5 inches

Father	Son
65.0	59.8
63.3	63.2
65.0	63.3
65.8	62.8
61.1	64.3
63.0	64.2
⋮	⋮

At this point everything we've done in linear regression has only been a mathematical result

- The Best-fit-line is a geometric minimization problem
- We have yet to make assumptions

Now, we will introduce the assumptions for SLR and then interpretations for the slope and intercept

- Assumptions are wrong -> best fit line is wrong
- ALWAYS check the assumptions before you worry about your interpretations/model's results
- NO INTERPRETATION for $\hat{\beta}_0$ or $\hat{\beta}_1$ is valid if the assumptions are broken (in a meaningful way)

Assumptions

Need to check all of the following (in any order)

1. X and y's relationship has been correctly identified with our model

2. The errors are normally distributed with mean 0

3. The errors are **independent and identically distributed** (iid)
   - independent: The errors are not correlated
   - identically distributed: The distribution of the errors is the same for any value of x
     - ★ Eg This includes **homoskedasticity** which is where the variance of residuals is constant
     - ★ Eg symmetry/skew should be roughly the same

The last of these can be abbreviated to

$$e_i \overset{\text{iid}}{\sim} N(0, \sigma) \tag{1}$$

Independence

This is the odd man out as it's hard to check visually....often achieved through randomization (random sampling or random assignment in experiments)

- Has to be checked via *critical reflection*
- Verrrry common assumption to mess up
  - *psuedo − replication* is when an observational unit is measured twice (or more) and treated as two (or more) units
  - Eg I weigh 2 students 5 times. I have 10 numbers but still only 2 students
  - Ways to deal with this (mixed models)
- Things to ask yourself:
  - Is there a reason one observation might influence another observation?
  - If I told you the errors of the observations around a given observation would you have any information?

Checking Assumptions

The majority of your assumptions can and will be checked visually using graphics created from the residuals. Two super common types

- Residual Plot
  - Also called the residual-by-predicted plot
  - x-axis is predicted value
  - y-axis is the residual
  - want a cloud of points centered around the residual = 0 line.
  - Super important graph
- QQ-plot (Q = Quantile)
  - x-axis is the predicted values
  - y-axis is the observed values
  - want a straight 45 degree line

COLOR and SHAPE your residuals by categorical variables!!!!

Identically Distributed

The goal is that regardless of where we are at on the x-axis the residuals look roughly the same.

- Examples include homoskedasticity
  - (next slide)
- Different symmetries along the residual $= 0$ line in the residual vs predicted graph
- Any a priori reason to believe that some observations/errors should behave differently than others that haven't been accounted for
  - Eg doing an experiment with green onions which you sourced from two different grocery stores

Homoskedasticity

Homo -> same
...skedasticity -> randomness

- Want the residuals to have an equal spread/st. dev.
- Do NOT want fan/trumpet shapes; things that bulge in the middle, etc...
- Often driven by underlying scientific mechanism
  - eg weight of infants has a narrow range compared to weight of childern compared to the weight of adults

Normality

Normal (bell-shaped) residuals are not critical for fitting the line but are important for inference about our line

- Want data points equally scattered above and below the line
- No noticeable pattern outside of a cloud of points
- Often violated by data that are proportions or who take counts near 0
  - Eg you can't have data outside of 0-10 for the number of heads in 10 coin flips
    - ★ Logistic regression is better
  - Eg Number of deer bagged during hunting season is very close to 0 but never negative and always a whole number
    - ★ Poisson regression is better

Relationship between X and Y

Our goal is to have our candidate model's line and our imaginary, free-hand drawn best fit line approximate each other well. If we pick a poor model it won't approximate the truth well.

- This can be checked if the residual plot has no noticeable pattern
- For a single explanatory variable this requires the scatterplot to be linear.
- Curves, exponential, logarithmic, periodic waves, etc... use other techniques
  - We will get into a few slide decks
- Can be checked from the scatterplot initially and again from the residual plot later on

Slope Interpretation

Est. regression equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$

What happens when we increase X by 1 unit?

Slope Interpretation

Est. regression equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$

What happens when we increase X by 1 unit?

We'd expect the average response variable to increase by the value $\hat{\beta}_1$, which is our slope

Slope Interpretation

The more formal interpretations of the slope $\hat{\beta}_1$ are....

Interpretation 1: For each 1 unit change in the explanatory variable (x), the mean value of the response variable (y) will change by the [value of slope], on average.

Interpretation 2: For each 1 unit change in the explanatory variable (x), the predicted mean value of the response variable (y) will change by [value of slope].

Intercept Interpretation

Est. regression equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$

What does $\hat{\beta}_0$ do? What's it's geometric interpretation?

Intercept Interpretation

Est. regression equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$

What does $\hat{\beta}_0$ do? What's it's geometric interpretation?

The intercept is value of our best fit line when $X = 0$. That is, it's where our line crosses the y-axis.

Intercept Interpretation

Est. regression equation: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$

Interpretation: When the explanatory variable (x) is zero, we predict the mean response variable (y) to have a value of [intercept value].

Ask yourself: Does the intercept interpretation make sense?

- Is the intercept value actually possible for our response variable?
- Is x = 0 a reasonable situation?
  - Temperature in Celcius? Yes
  - Weight of a car? No

Interpretation Tips

- Always take a deep breath and ask yourself if your regression line makes sense and do you see why it is where it is
- Interpretations for $\hat{\beta}_0$ are rarely meaningful/physically possible
- Our interpretations deal with *generalities*
  - Always want to say something like "we believe/predict" or "on average"
  - Don't sound definitive
    - ★ eg "Our salaries in 10 years will increase if Prof Paris gets a sizable raise" is incorrect
    - ★ Usually this ties into correlation not being causation

Pearson's Height Data – Interpretations

$$\widehat{\text{Son's Height}} = 33.9 + 0.51 \times \text{Father's Height}$$

Slope Interpretation:

For each 1 inch change in Father's height, the prediction for son's height changes by 0.51 inches. OR

For each 1 inch change in Father's height, the son's height changes by 0.51 inches, on average, we believe.

Pearson's Height Data – Interpretations

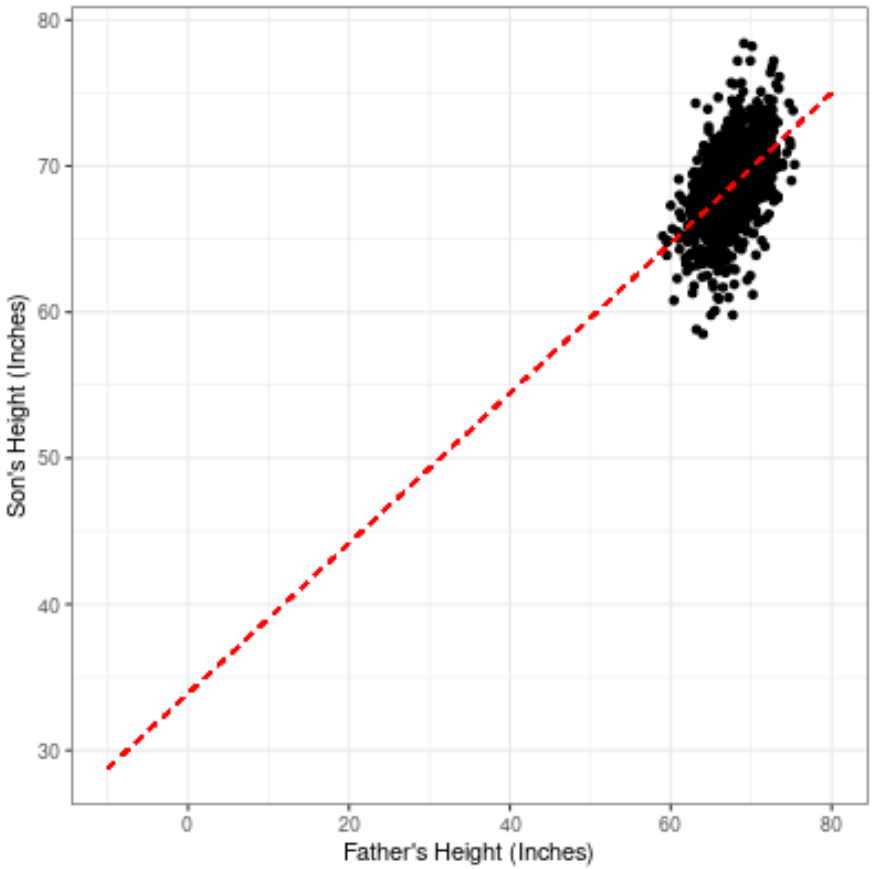
$$\widehat{\text{Son's Height}} = 33.9 + 0.51 \times \text{Father's Height}$$

Intercept Interpretation:

When the father's height is zero inches, the predicted height for the son is 33.9 inches.

- does this make sense?

Intercept and Extrapolation



Extrapolation means making predictions for values outside the area of our data

- These predictions are unreliable, since we don't know if the relationship is true for these values

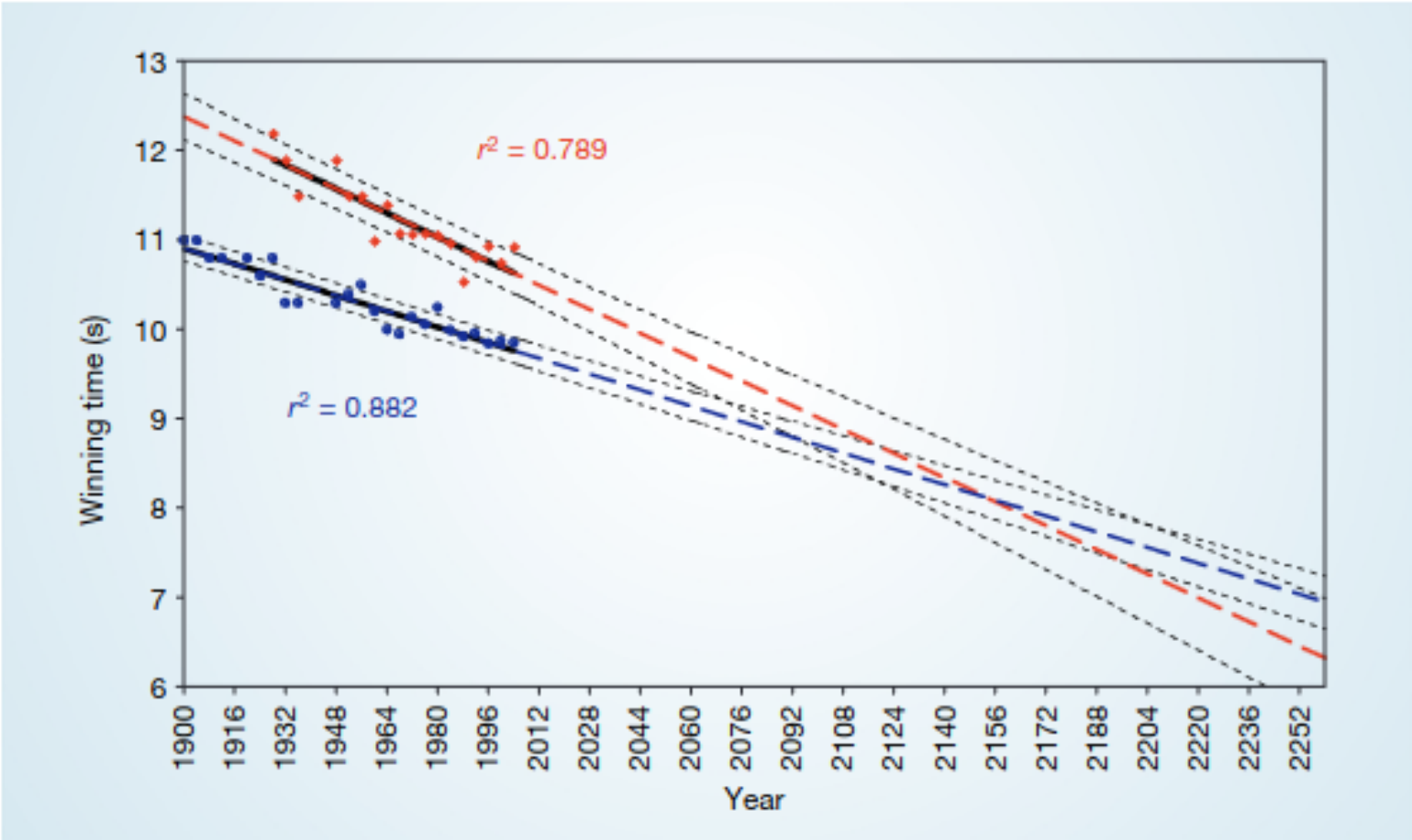
Extrapolation

In 2004, an article was published in *Nature* titled "Momentous sprint at the 2156 Olympics." The authors plotted the winning times of men's and women's 100m dash in every Olympic contest, fitting separate regression lines to each; they found that the two lines will intersect at the 2156 Olympics. Here are a few of the headlines:

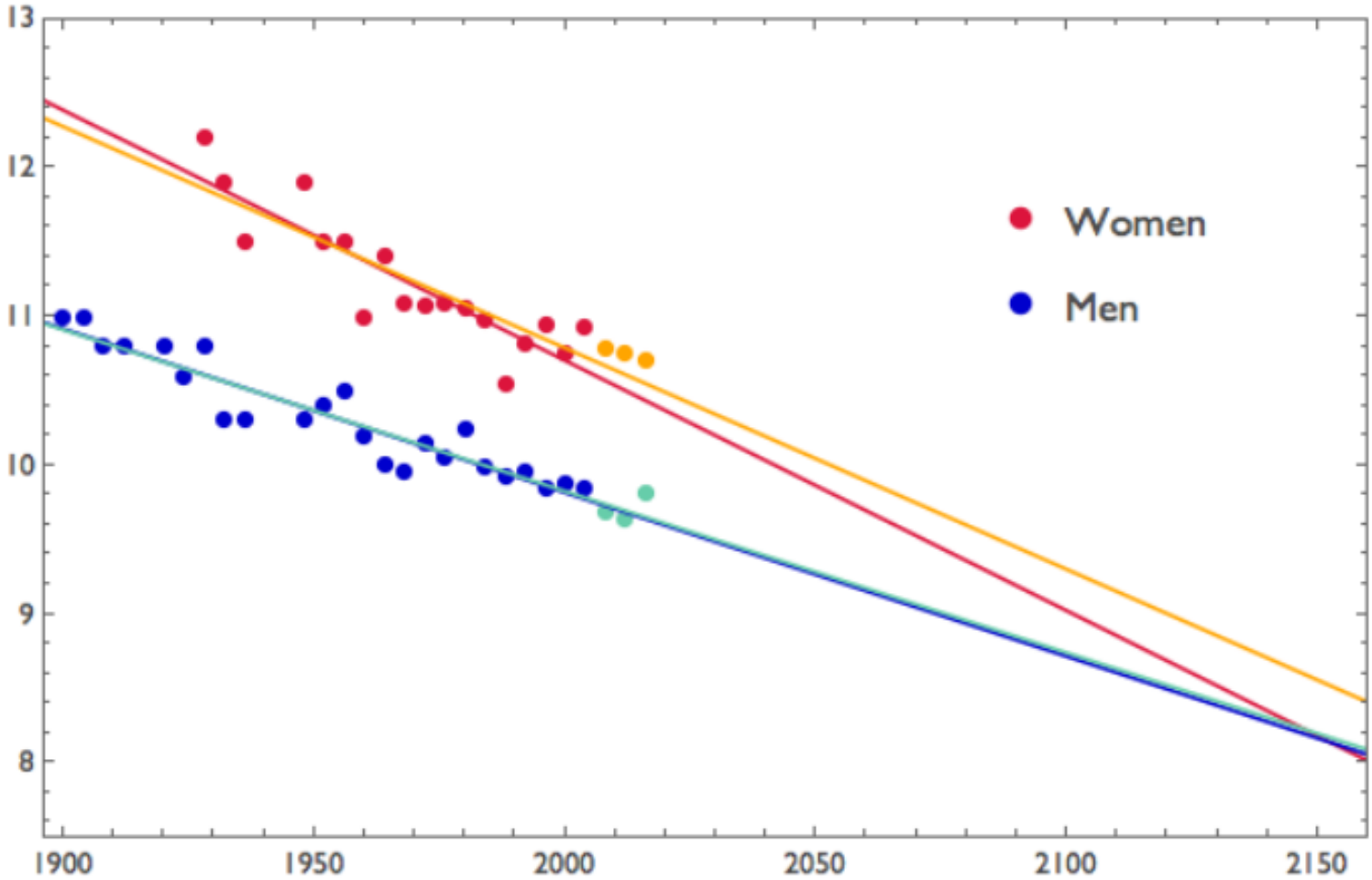
- "Women 'may outsprint men by 2156'" – BBC News
- "Data Trends Suggest Women Will Outrun Men in 2156" – Scientific American
- "Women athletes will one day out-sprint men" – The Telegraph
- "Why women could be faster than men within 150 years" – The Guardian

Momentous sprint at the 2156 Olympics?

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

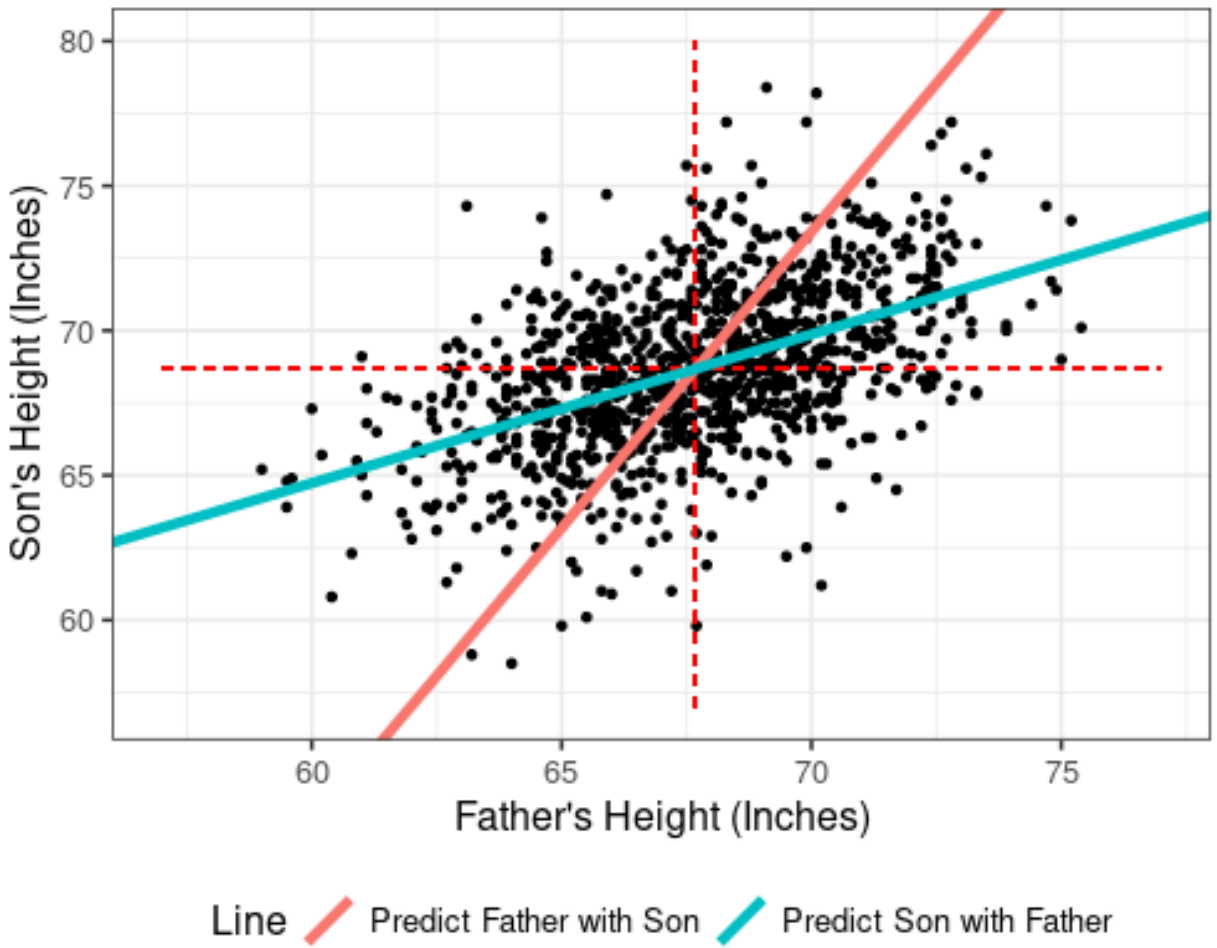


12 years of data later



Asymmetry

Unlike correlation regression is *asymmetrical*: the choice of explanatory and response variables matter for the line



Sum of Squares

There exists a geometric relationship

The total sum of squares (TSS, variability) in y can be broken down into two pieces. The sum of the square of the errors (SSE) and the sum of the squares of the model (SSM).

$$\text{TSS} = \text{SSM} + \text{SSE}$$

R^2 is SSM / TSS

That is, it's how much variability in y we can explain using our model

Assessing Quality of Fit

Coefficient of determination (R^2)

- measures how close the observations match the predictions
- ratio written as decimal or percentage between 0% and 100%
- larger values imply better fit, stronger linear relationship between the variables
- It's a single statistic and can be useless at times

Interpretation:
R^2 is the percentage of variation in the observed values of the response variable (y) that can be explained with the linear regression model including the explanatory variable (x). [include context]

Assessing Quality of Fit

We also saw that the **correlation coefficient (r)** can be used to quantify the strength of the linear relationship.

There is a connection between r and R^2.

- $r^2 = R^2$
- $\text{r} = \pm\sqrt{R^2}$ (need to find the correct sign using scatterplot / slope)

Yes, the relationship really is that simple, R^2 is r squared