Math 183
Statistical Methods

Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today: Chapter 7 (beginning)

- Learn to describe a scatterplot using FDSO
- Get a visual and intuitive sense of the correlation coefficient R
- Residuals
- Least square method for regression

More Inference for Two Numerical Variables

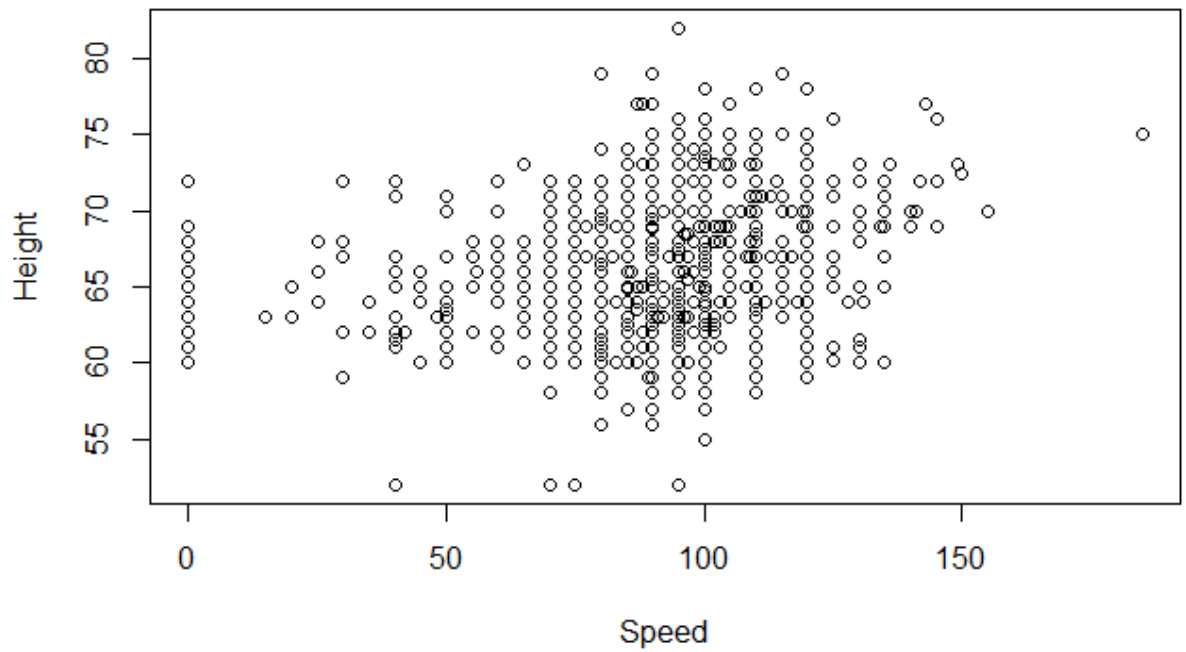
Two categorical variables:

- Display via contingency table
- Inference via χ^2

Two quantitative variables:

- Display via scatterplot
- Inference depends if data are about 1 population or 2 and what you're exploring.

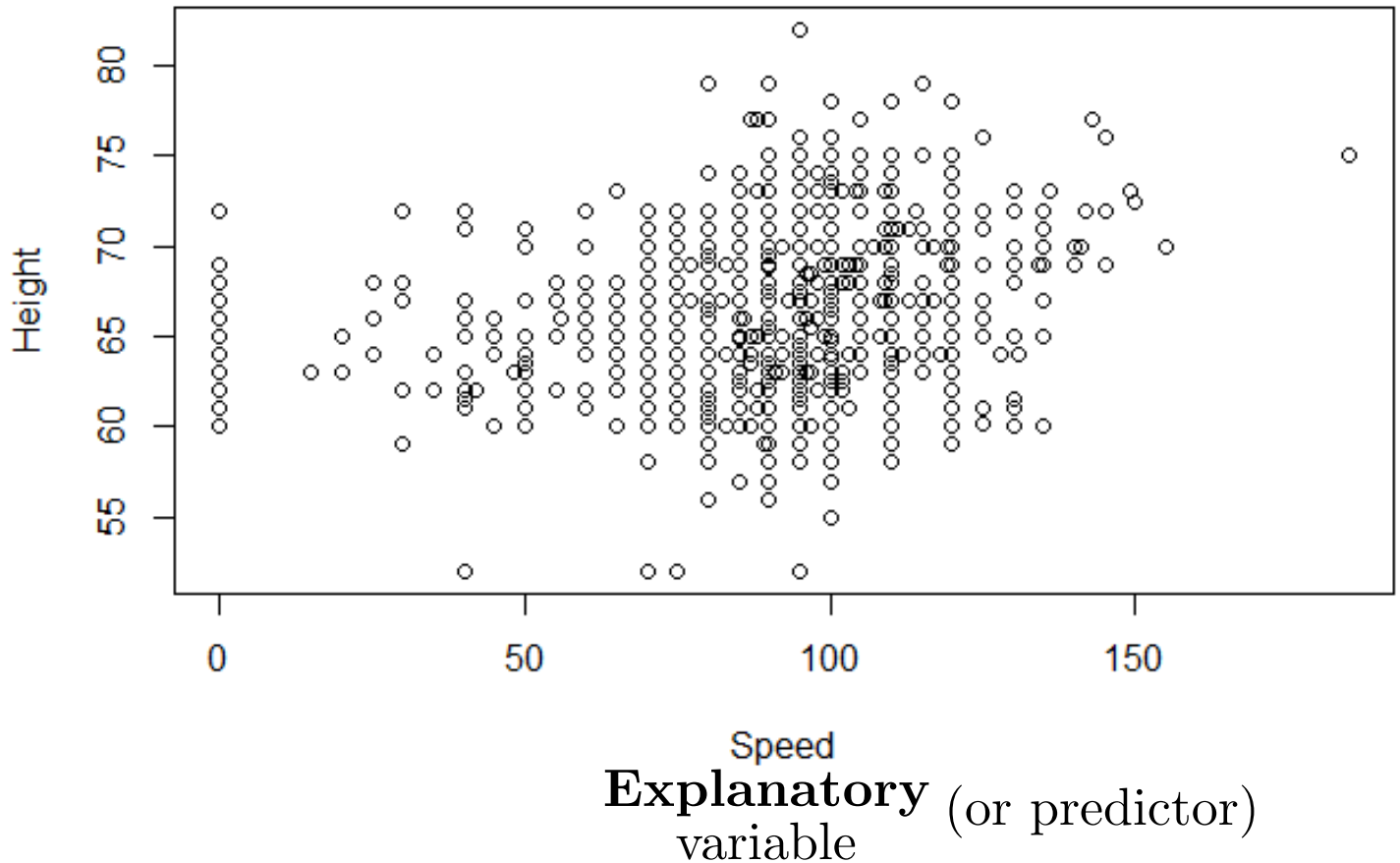
Speed	Height
85	69.0
40	71.0
87	64.0
110	60.0
110	70.0
120	61.0
90	65.0
90	65.0
80	61.0
95	69.0
110	63.0
90	72.0
110	70.0
70	68.0



Each row holds two pieces of data about the **same** person.

Response and Explanatory Variables

Response variable



Explanatory variable (or predictor)

A **Scatterplot** shows the relationship between two <u>numeric</u> variables.

How to Discus Scatterplots: FDSO

Form:

Linear Curve Repeating/Oscillating

Direction:

Positive Negative No direction

How to Discus Scatterplots: FDSO

Strength:

Very strong Strong Weak

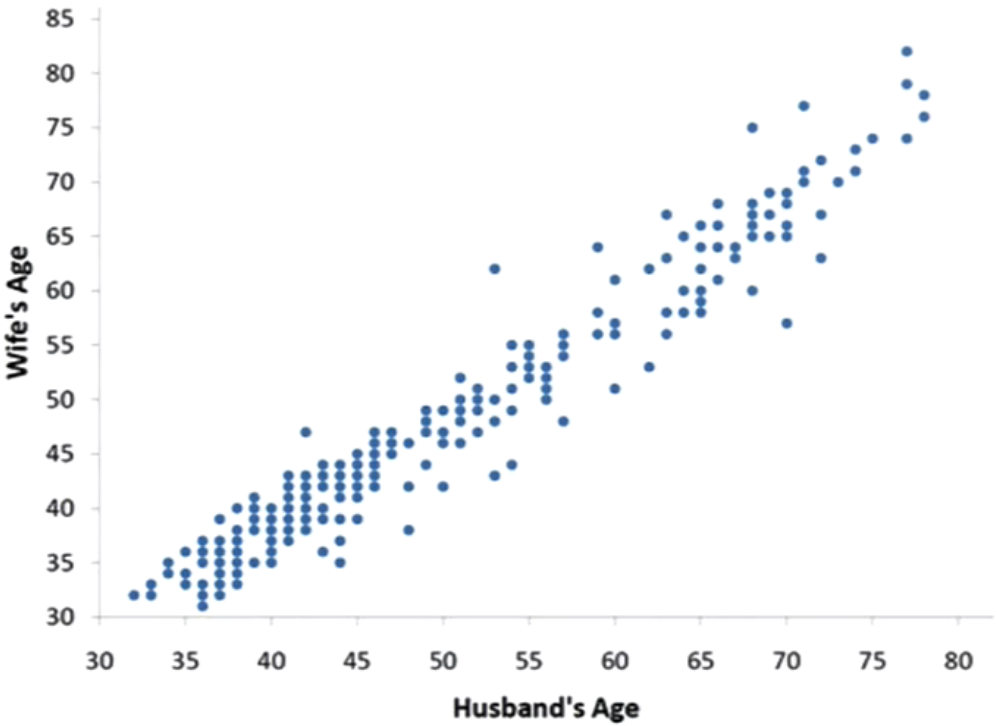
Outliers:

This scatterplot has a weak to moderate, positive, linear association. We have an outlier.

As with a single quantitative variable, the notion of outlier is vague. Just think: "A point that stands far apart from the overall trend in the scatterplot."

Your Turn! Describing Real Data Using FDSO

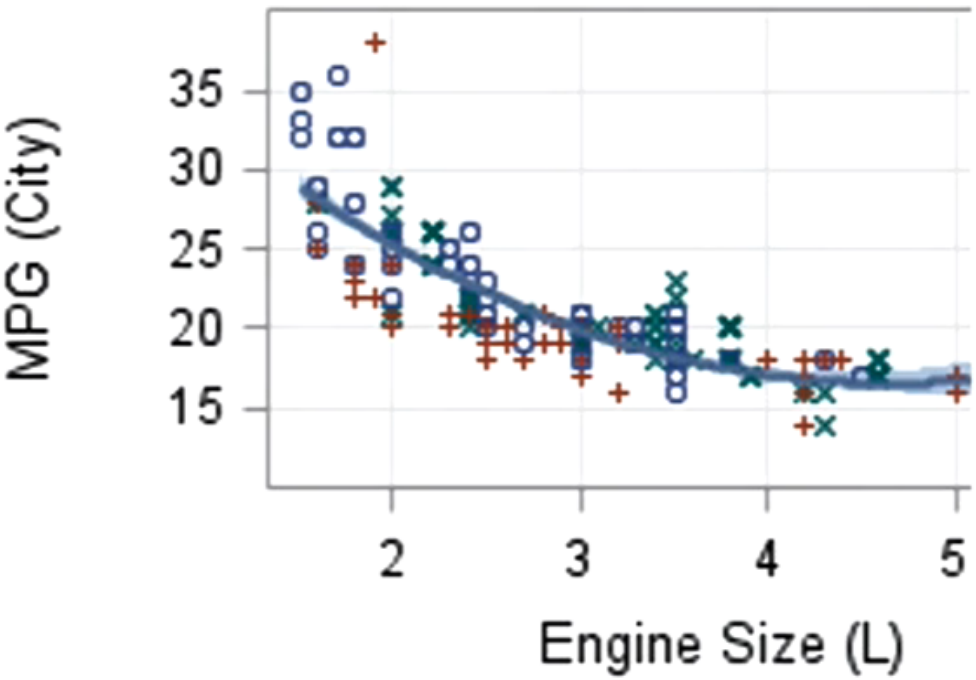
Age of husband and Age of wife



- Form: Linear
- Direction: Positive
- Strenght: Strong
- Outliers: None

Your Turn! Describing Real Data Using FDSO

Engine size of a car (liters) and Fuel economy (MPG)



- Form: Curved
- Direction: Negative
- Strenght: Moderate
- Outliers: Maybe the top red plus...?

Moving Beyond Vague Language

Correlation: A statistic that measures the Srength and Direction of a linear association (Form) between two quantitative variables where no Outliers are present.

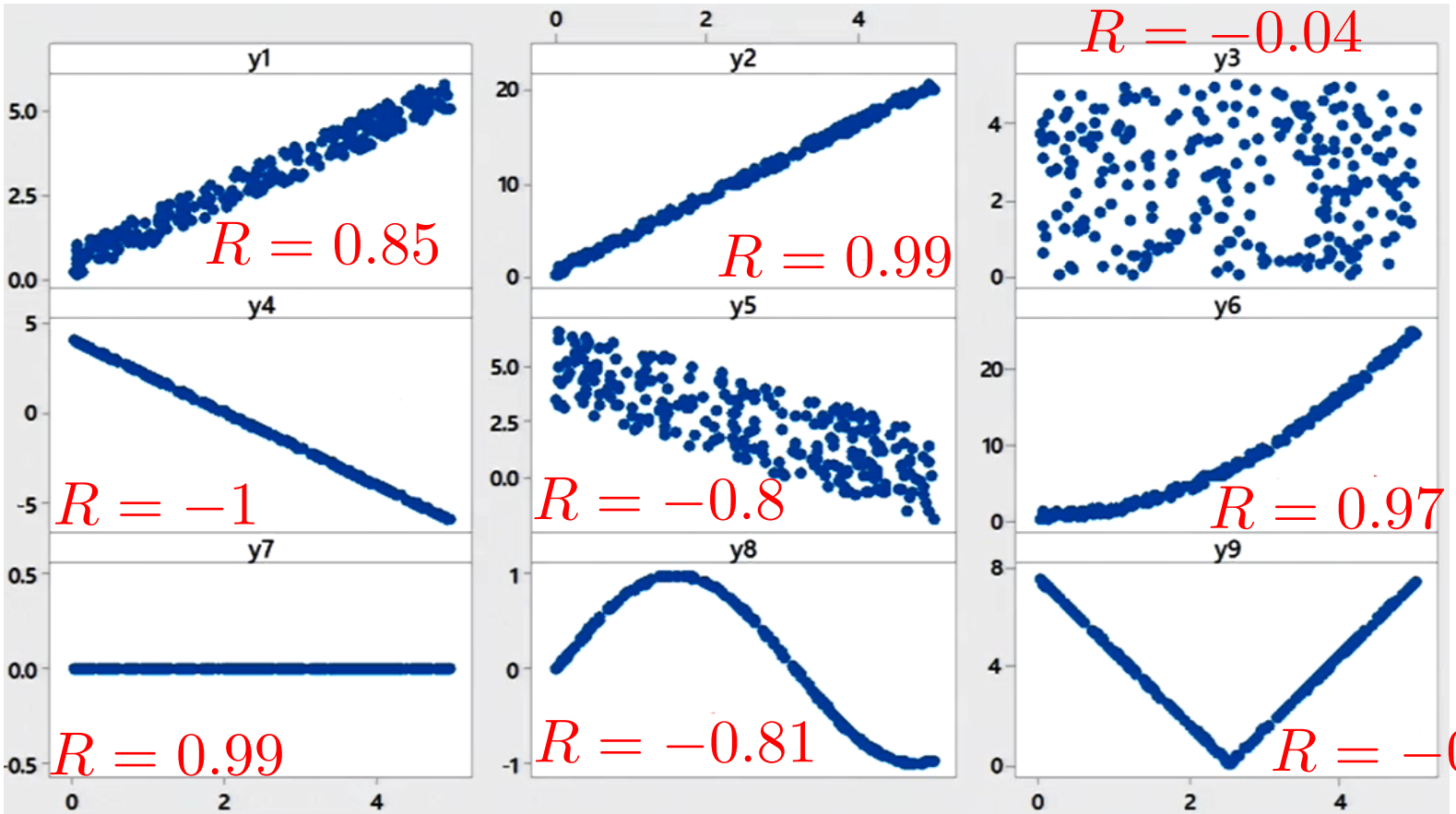
Reported using either an uppercase R or a lowercase r:

$$R = r = -0.3.$$

Note: The programming language R is NOT named after the R statistic, but for the first names of its inventers (Ross Ihaka and Robert Gentleman) and as a reference to a related language S.

Deriving Knowledge about R Using Examples

Looking at these examples, make observations about the R statistic.



Facts About R, The Correlation Statistic

- We always have $-1 \leq R \leq 1$.
- The Direction is encoded in the sign:

$$\begin{aligned} + &: \quad \text{positive association} \\ - &: \quad \text{negative association} \end{aligned}$$

- The Srength is encoded in the value:

$$\begin{aligned} \text{Stronger association} &: \quad \text{occurs when } R \text{ is closer to 1 or -1} \\ \text{Weaker association} &: \quad \text{occurs when } R \text{ is closer to 0} \end{aligned}$$

- 1 and -1 are possible: the data must perfectly lie on a line
- Choice of predictor/response doesn't matter:

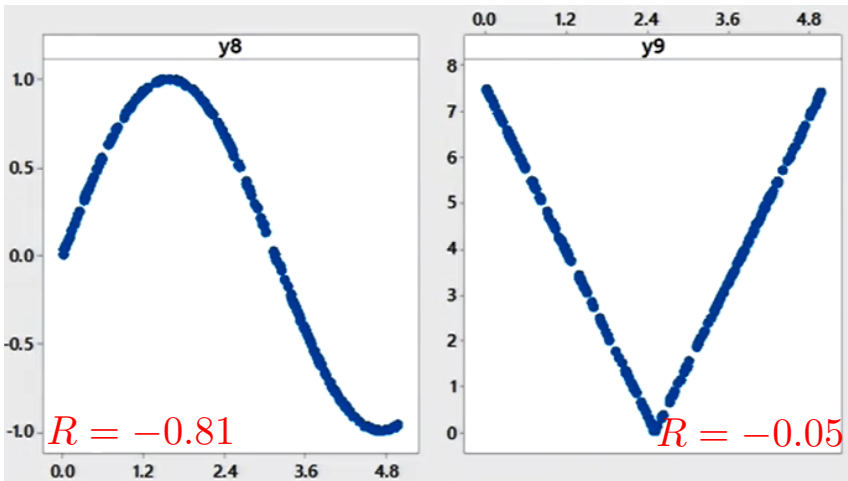
$$cor(X, Y) = cor(Y, X)$$

- Correlation is a unitless idea
- Correlation is unaffected by linear scale changes

$$cor(X, Y) = cor(X, 3.14Y) = cor(X, Y + 100) = cor(2X - 6.8, 95Y + 7)$$

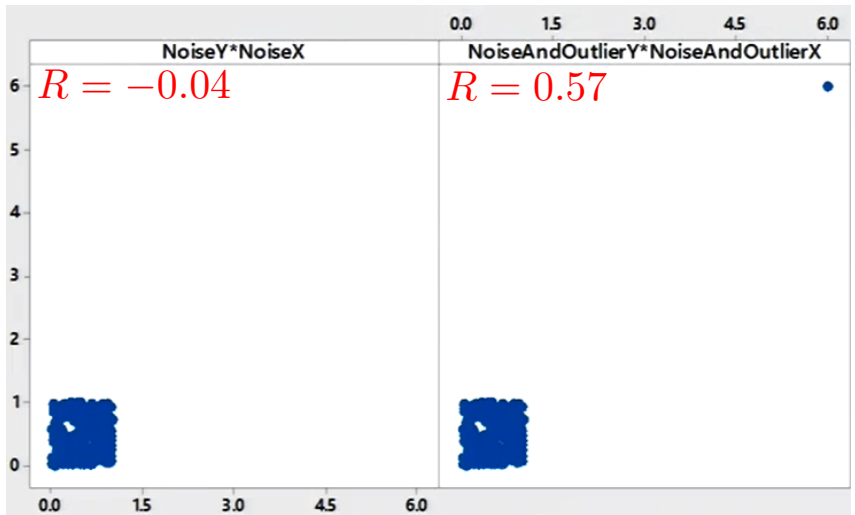
Cautions About R

Never use R to measure correlation for associations that are non-linear.



Correlation doesn't measure strength in non-linear associations

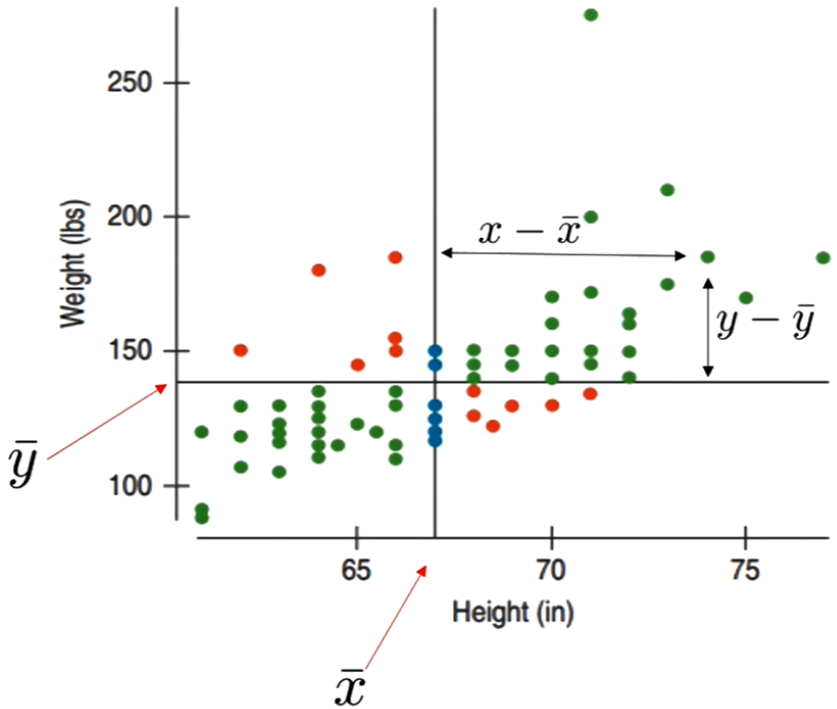
Never use R to measure correlation if outliers are present (even one!).



The effect of a single outlier can be huge.

How is R Calculated?

$$R = \frac{1}{n-1} \sum_{(x,y) \text{ pairs}} \frac{(x-\bar{x})}{s_x} \frac{(y-\bar{y})}{s_y}$$



Numerator: Factors that accumulate positive and negative directions based on the individual points

Denominator: Factors that eliminate scaling for each axis, and ensure that $-1 \leq R \leq 1$.

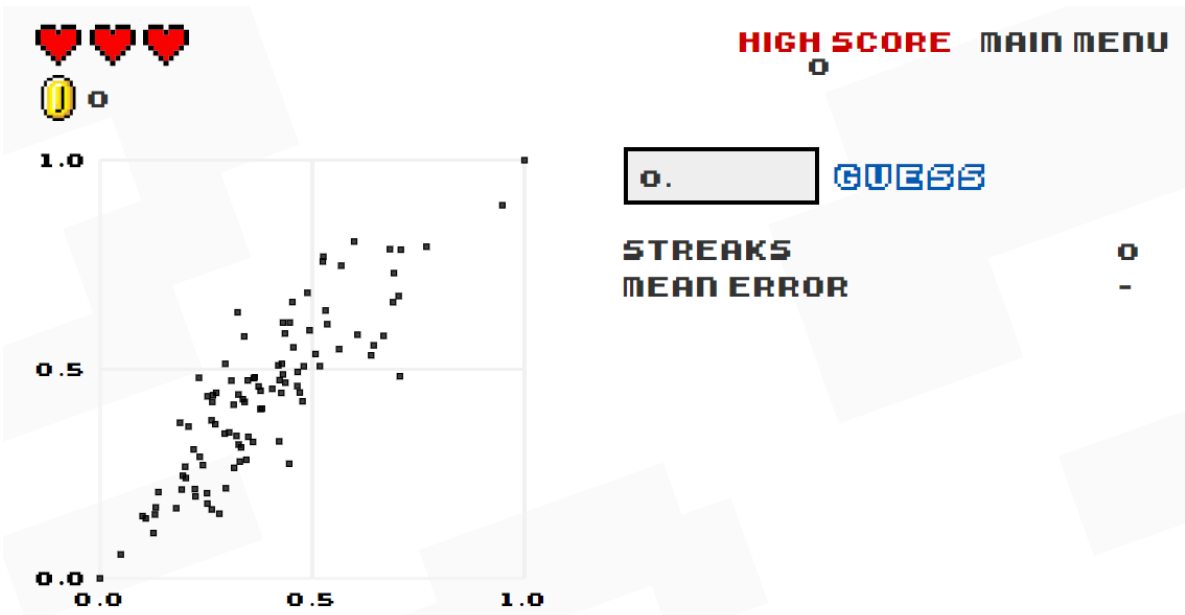
(R is almost always calculated in a statistical program, not by hand)

Your Turn!

On your cell phone, go to

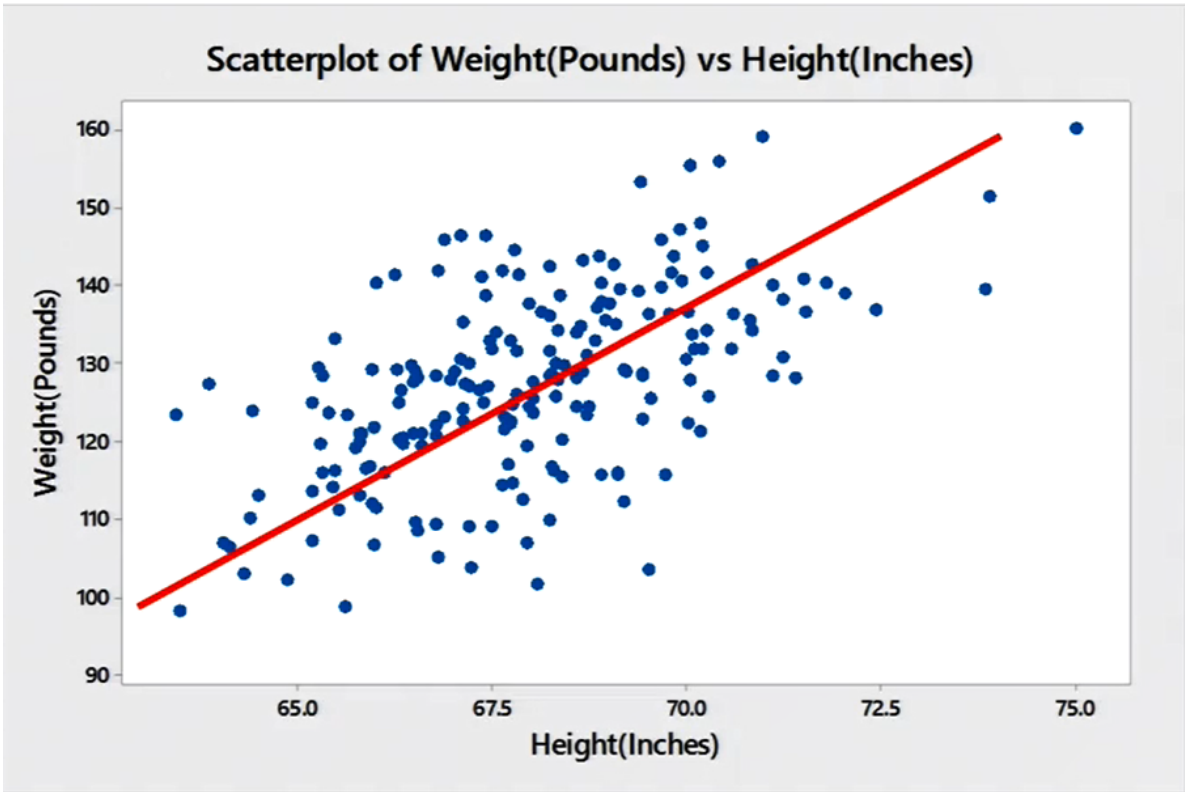
http://guessthecorrelation.com/

and play for one minute.



Moving Beyond Correlation to Prediction

If we had to draw a straight line through our scatterplot that "fit the data the best", it might look like the red line.



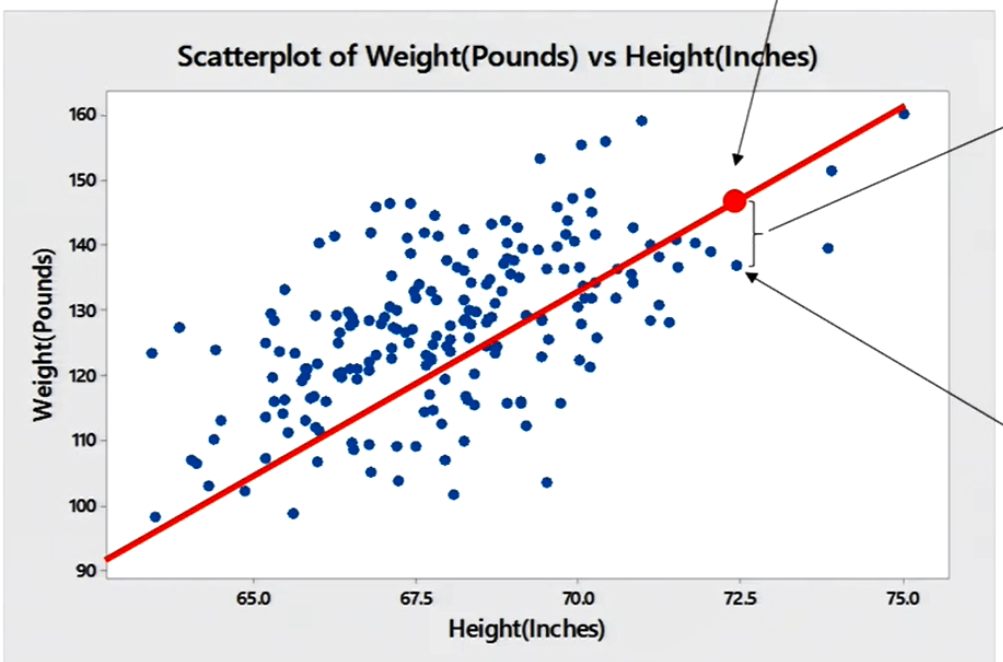
A **linear model** is an equation the "best fits" the data. It is also as the "line of best fit", the "regression line", and the "least squares line".

Observed Value, Predicted Value, Residual

Given any point (x, y) in a scatterplot, we now have two key ideas: for x fixed,

- y: the **value observed** from actual data point
- $\hat{y}$: the **value predicted** from model

From these values, we can calculate the **residual** $e = y - \hat{y}$, which measure how off the model is at the value x.



$(x, \hat{y}) = (72.4, 143)$

Residual at $x = 72.4$

$e = 136 - 143 = -7$

$(x, y) = (72.4, 136)$

What is the "Best Fit"?

Ideally, all the residuals would be 0 (a perfect model!).

Decause data are rarely perfect, we could define

$$\text{model error} = \sum (\text{residuals for each data point})?$$

Then, the line of best fit makes this expression minimal.

Two choices to fix this issue are:

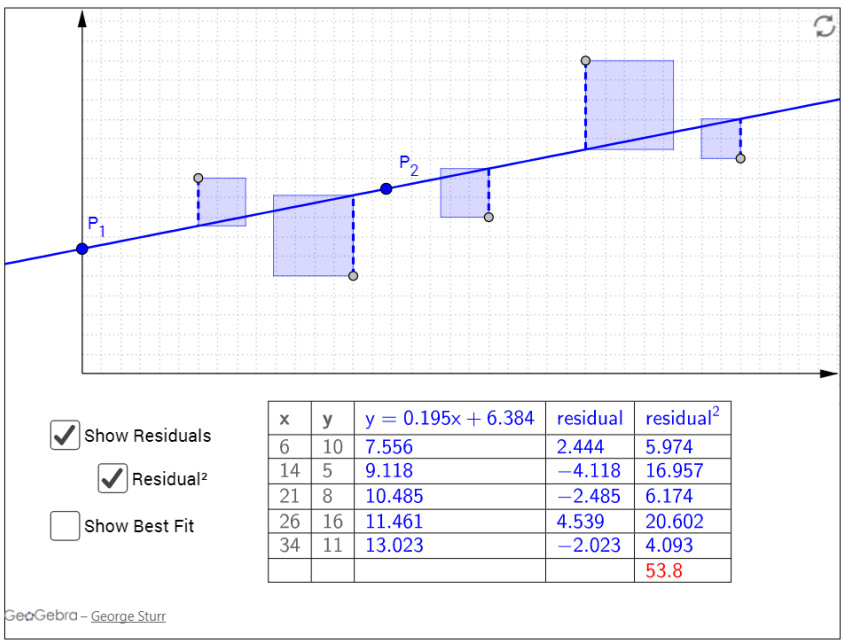
$$\text{model error} = \sum |\text{residuals}| \quad \text{or} \quad \text{model error} = \sum (\text{residuals})^2 .$$

The second choice is standard, penalizes large residuals, and is easily differentiable.

Exploring Linear Regression Visually

Try out this interactive GeoGebra tool:

https://www.geogebra.org/m/dlsxY1uX



x	y	y = 0.195x + 6.384	residual	residual²
6	10	7.556	2.444	5.974
14	5	9.118	−4.118	16.957
21	8	10.485	−2.485	6.174
26	16	11.461	4.539	20.602
34	11	13.023	−2.023	4.093
				53.8

- Move p_1 and p_2 until you *feel* you have the line of best fit. Check your guess by showing the line of best fit.
- Refresh the page. Try checking "*Show Residuals*" and "*Residuals*^{2}", and then minimize the sum of squared residuals (red number). Then check your guess.