CORRELATION & REGRESSION

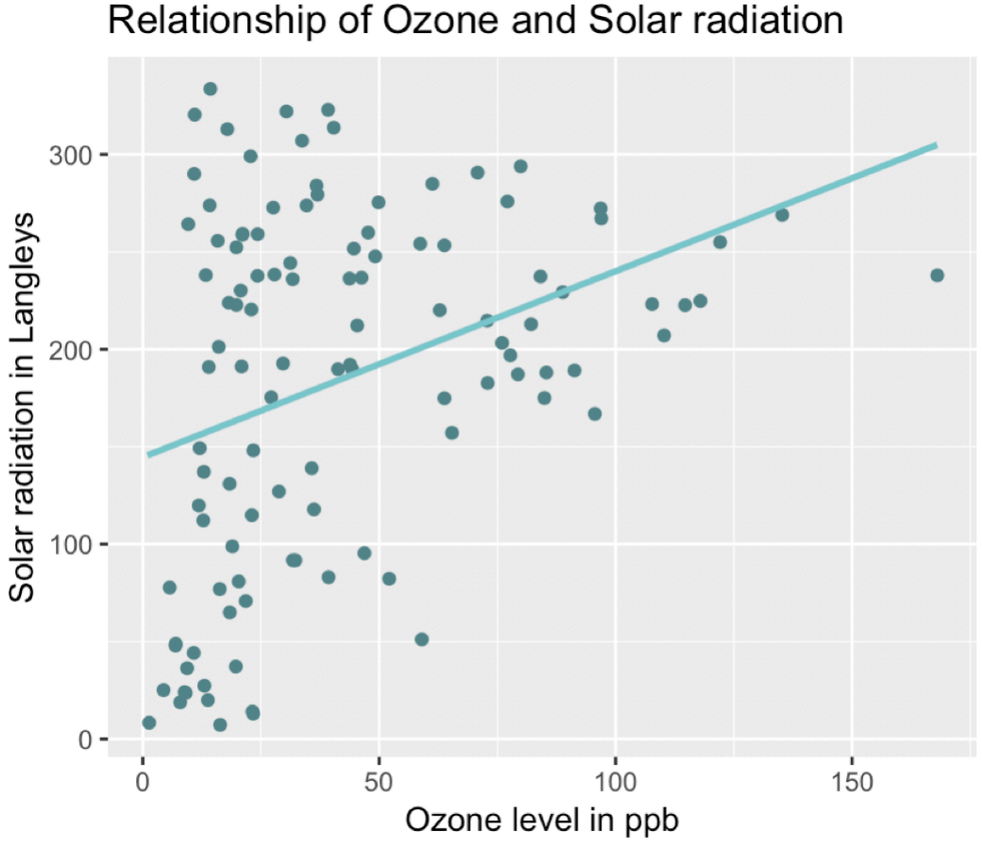
Tutorial #5

A RECAP ON P-VALUES

- MrNystrom - What is a P Value? What does it tell us? https://www.youtube.com/watch?v=-MKT3yLDkqk

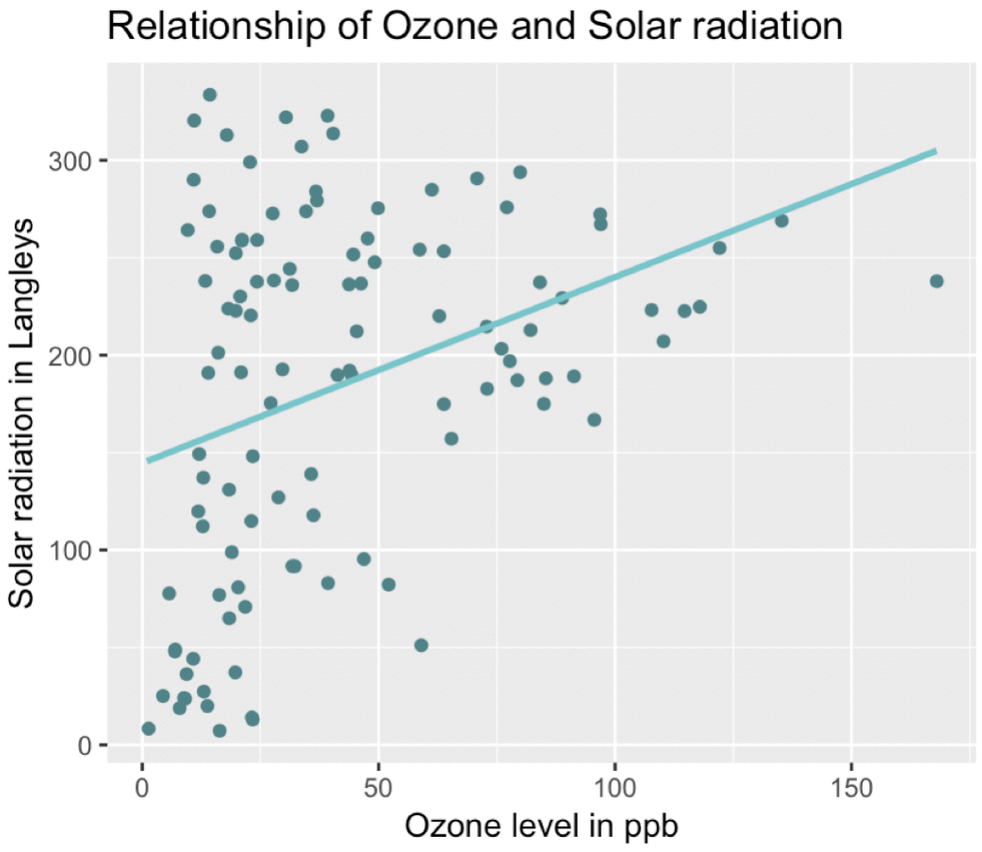
THE CORRELATION

- Describes the relationship of two variables
- can be written as: y~x (y *is explained by* x)
- requires (ordinal)/interval/ratio data
- **result**: the so-called correlation coefficient, can be between -1 and 1



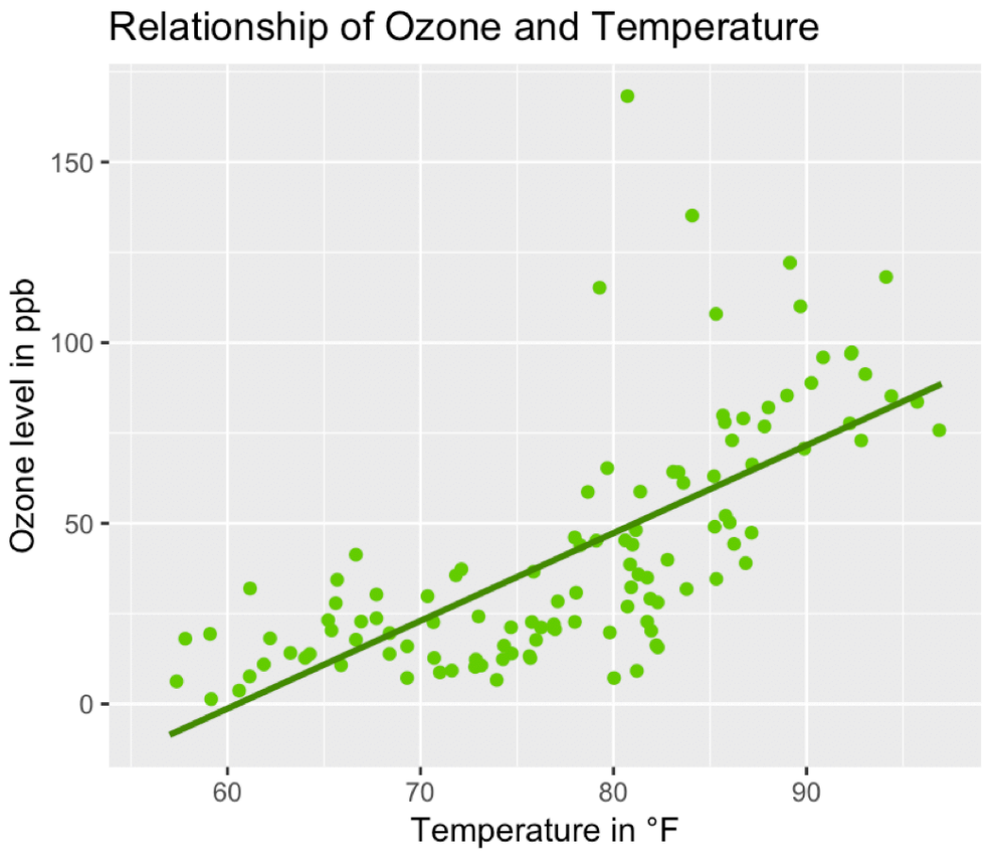
SOME CORRELATION VOCABULARY

A plot with lots of points is called a scatterplot.



0.35

weak correlation

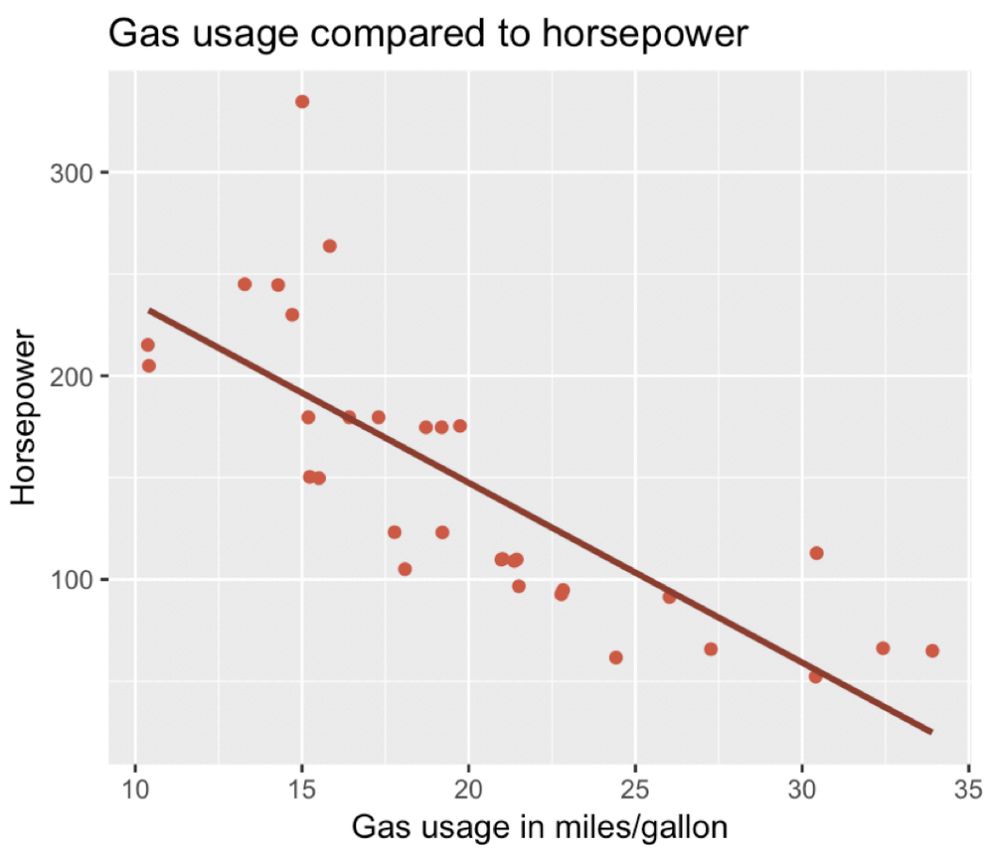


0.7

strong correlation

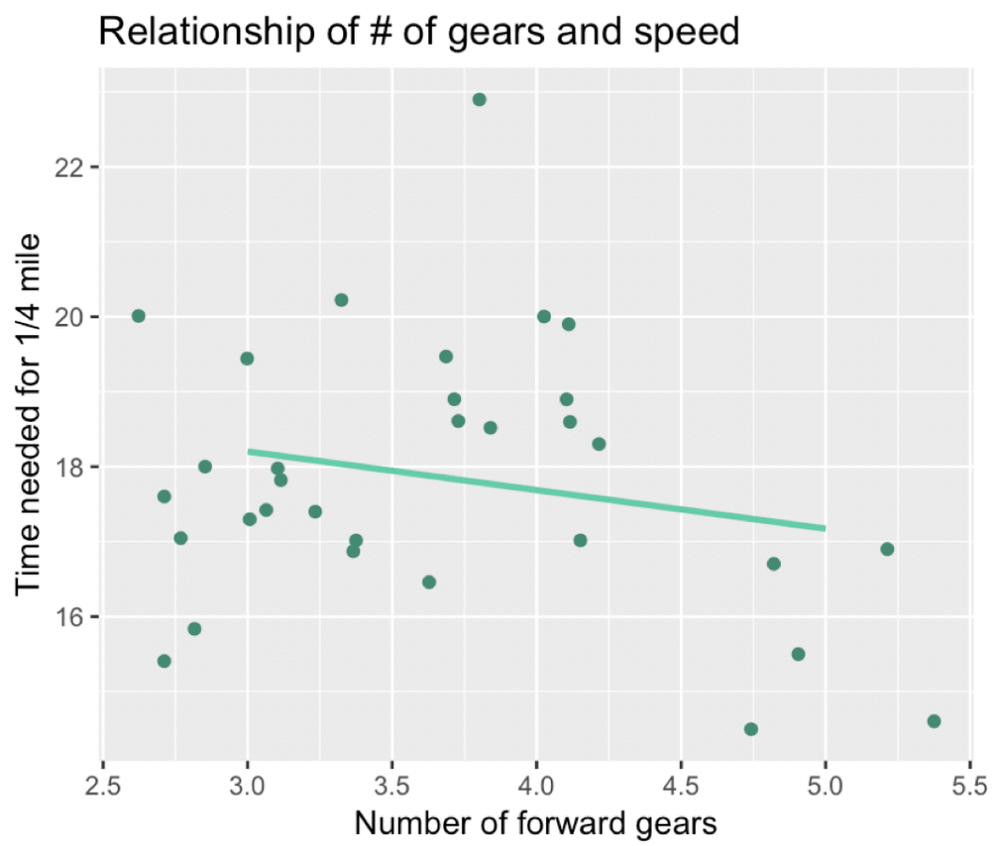
*Interpretation sentence: the **higher** the temperature the **higher** the ozone level.*

SOME CORRELATION VOCABULARY



-0.78

strong negative correlation

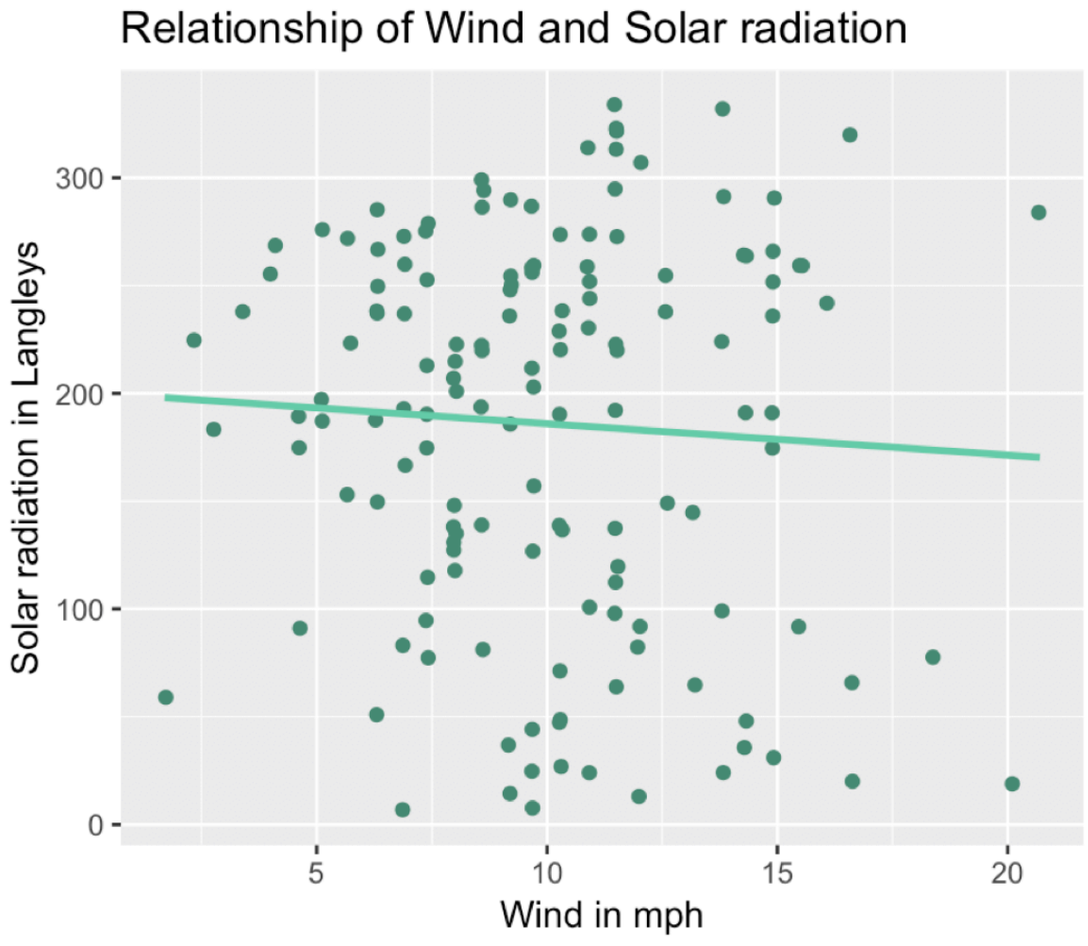


-0.21

weak negative correlation

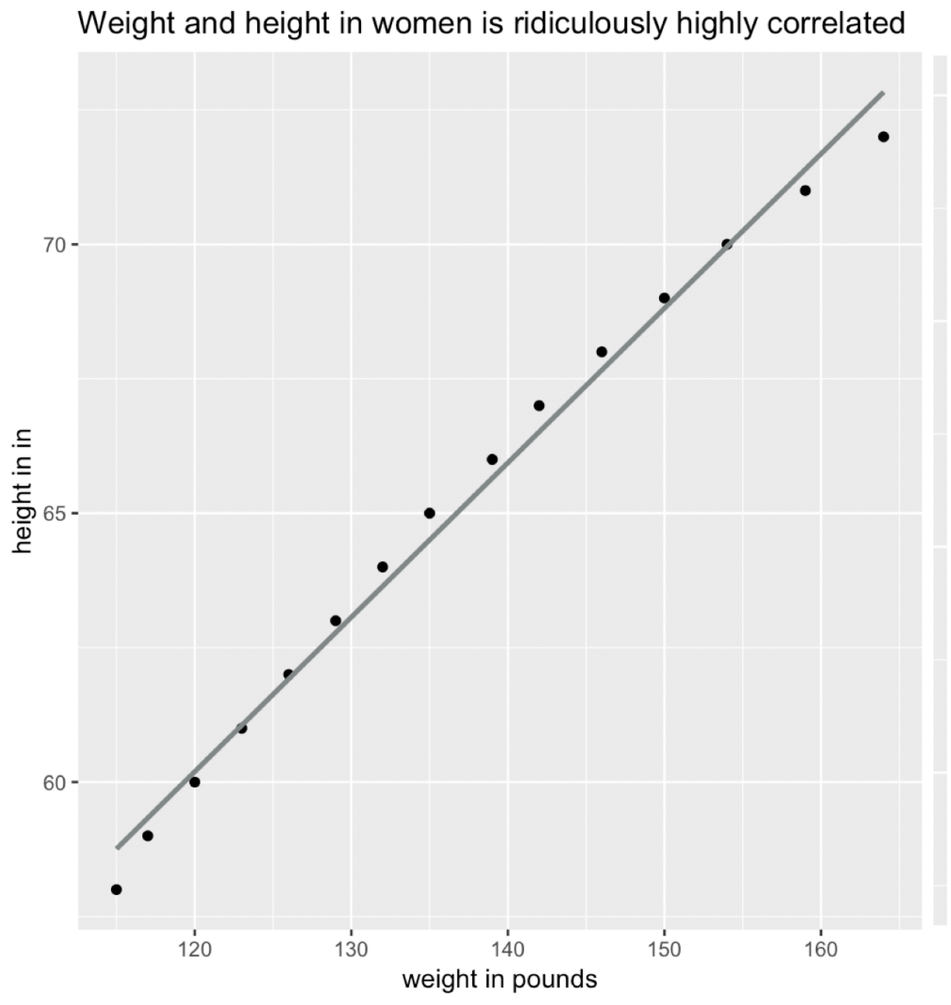
*Interpretation sentence: the **higher** the gas consumption the **lower** the horsepowers.*

WHAT WOULD MY DATA LOOK LIKE IF I HAD R=1 OR R=0?



-0.06

not correlated

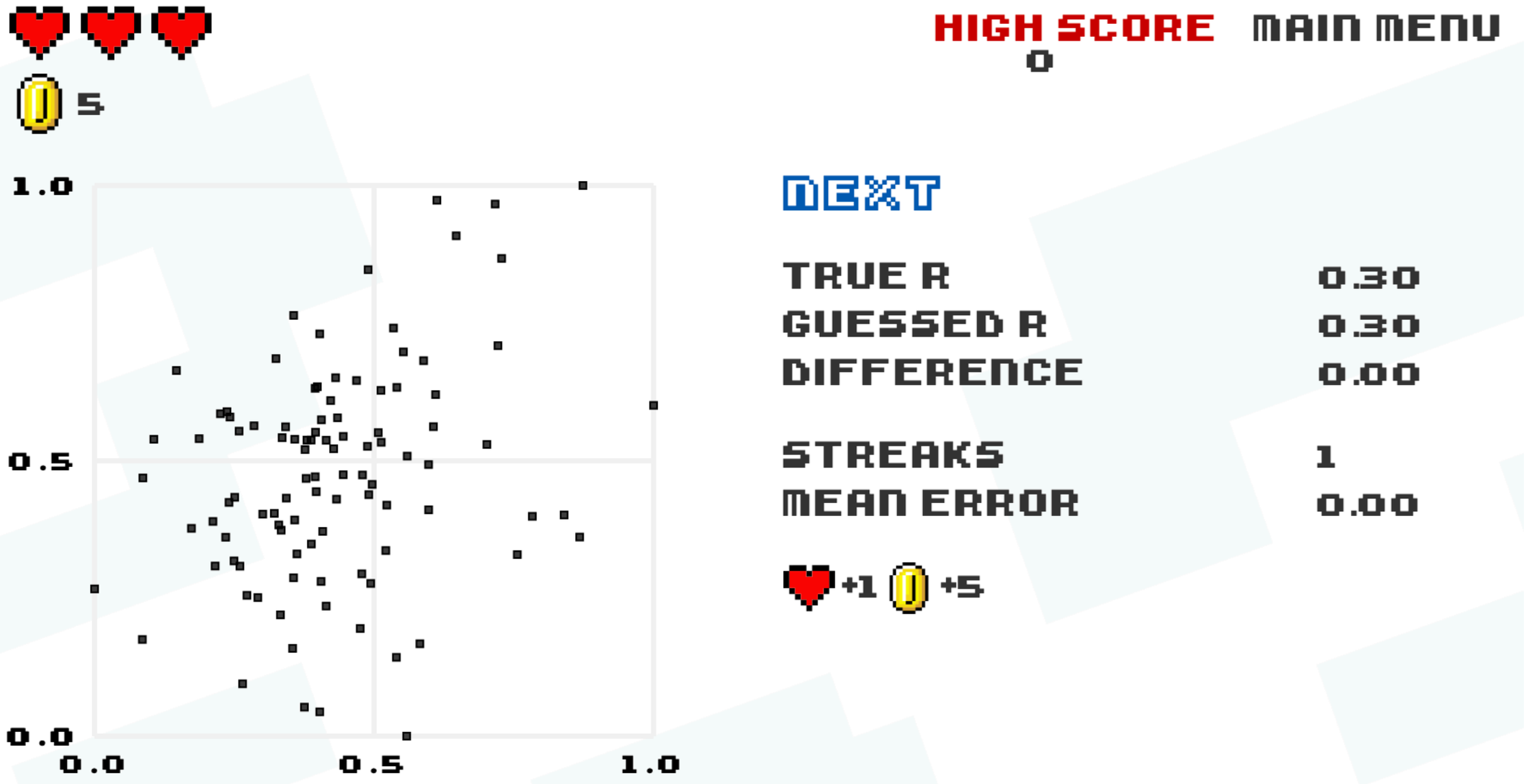


0.995

very strong correlation

LET'S PLAY A GAME!

- guessthecorrelation.com/

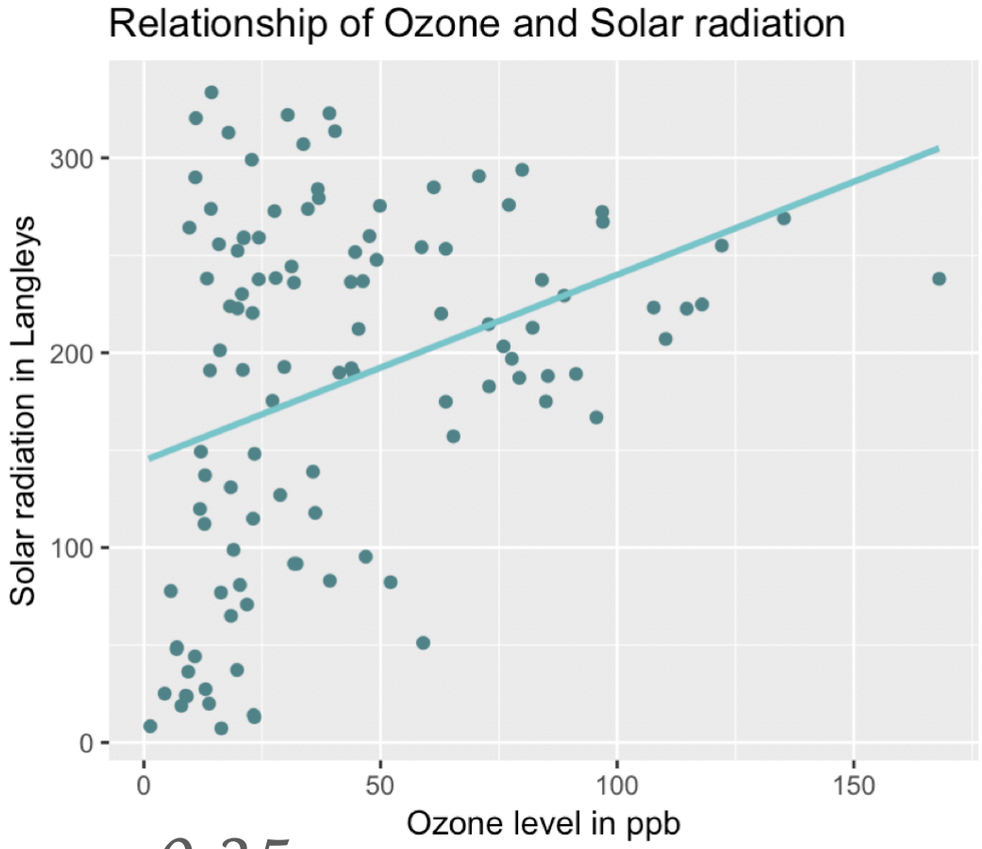


TESTING YOUR CORRELATION

- Sometimes, you might have too much variance or too few data
- Sometimes, your correlation can be due to some randomness
- That's why we test whether a correlation is significant.

Test	Data type	Purpose	Null Hypothesis	Alternative Hypothesis
t-test	Interval/Ratio	2 samples significantly **different**?	The two samples means are not different.	The two samples means are different.
correlation test	Interval/Ratio	Is there a correlation?	No correlation.	There is a correlation.

CORRELATIONS IN R



0.35

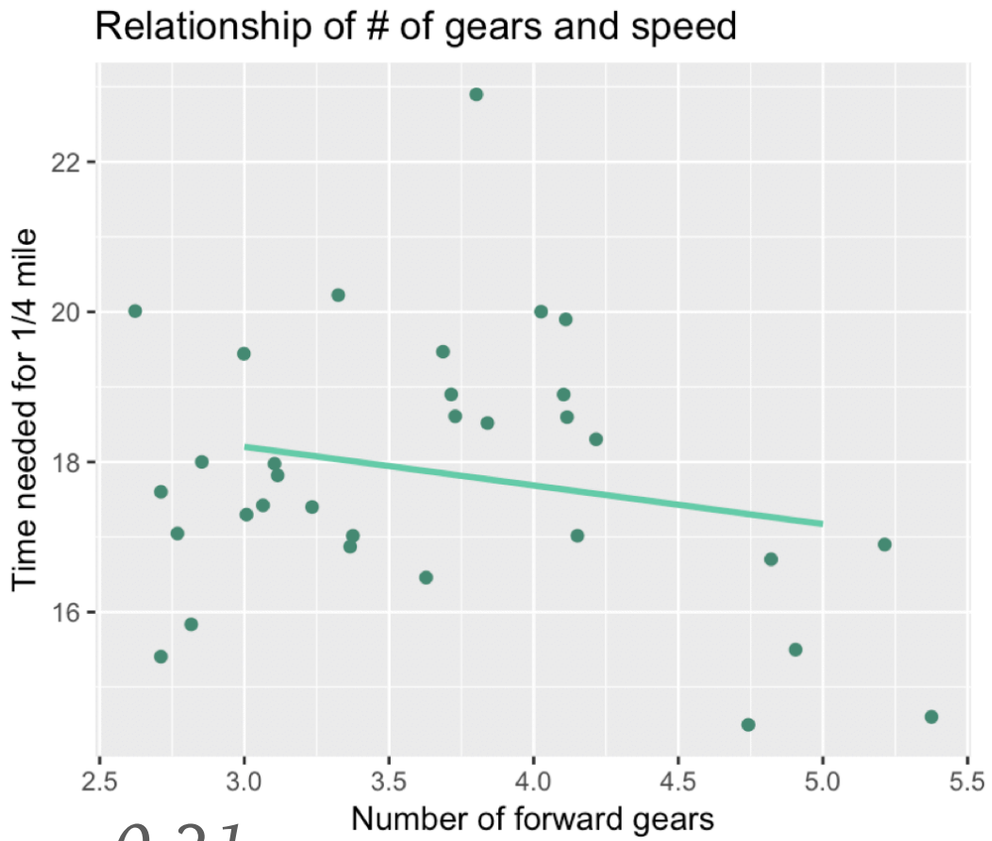
weak correlation

```
> cor(airquality$Ozone, airquality$Solar.R)
[1] 0.3483417
> cor.test(airquality$Ozone, airquality$Solar.R)

        Pearson's product-moment correlation

data:  airquality$Ozone and airquality$Solar.R
t = 3.8798, df = 109, p-value = 0.0001793
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.173194 0.502132
sample estimates:
      cor
0.3483417
```

H0: Correlation is equal to 0.

H1: Correlation is not equal to 0 (= my variables are correlated).

CORRELATIONS IN R



-0.21

weak negative correlation

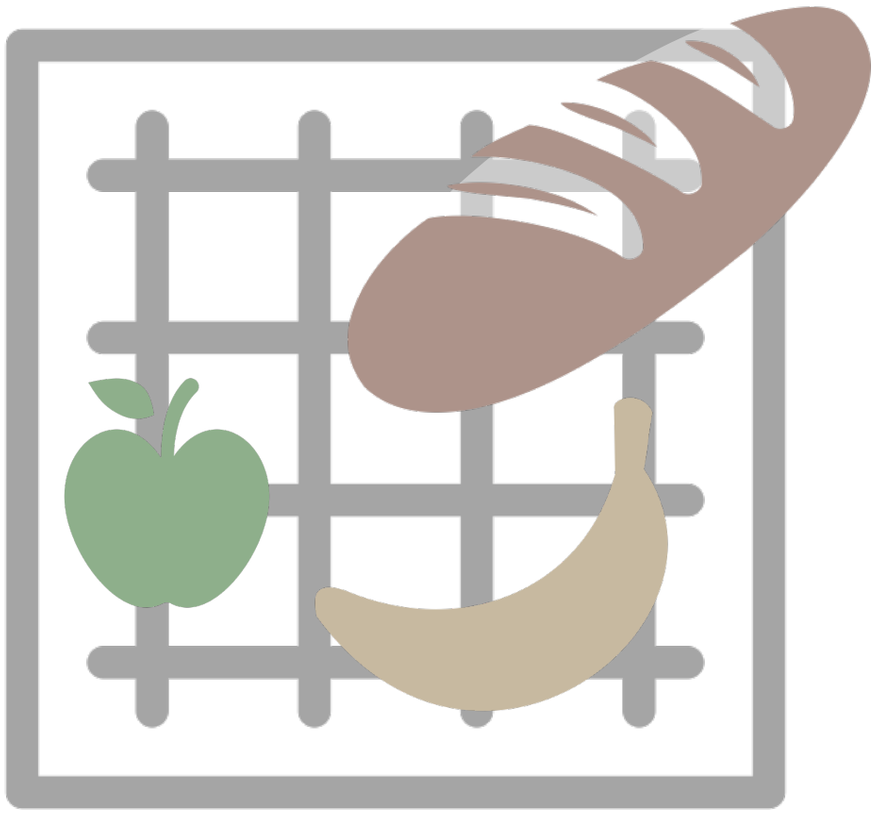
```
> cor(mtcars$gear, mtcars$qsec)
[1] -0.2126822
> cor.test(mtcars$gear, mtcars$qsec)

        Pearson's product-moment correlation

data:  mtcars$gear and mtcars$qsec
t = -1.1922, df = 30, p-value = 0.2425
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5226183  0.1469065
sample estimates:
       cor
-0.2126822
```

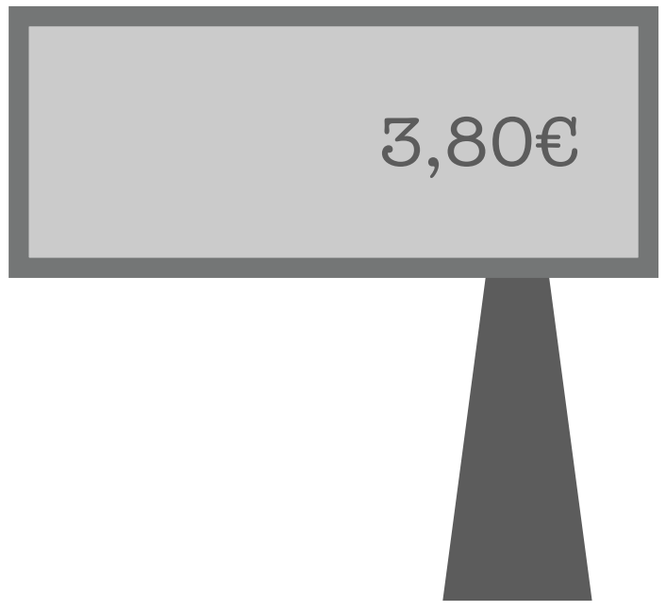
H0: Correlation is equal to 0.

H1: Correlation is not equal to 0 (= my variables are correlated).

REGRESSION

WHAT'S A MODEL?



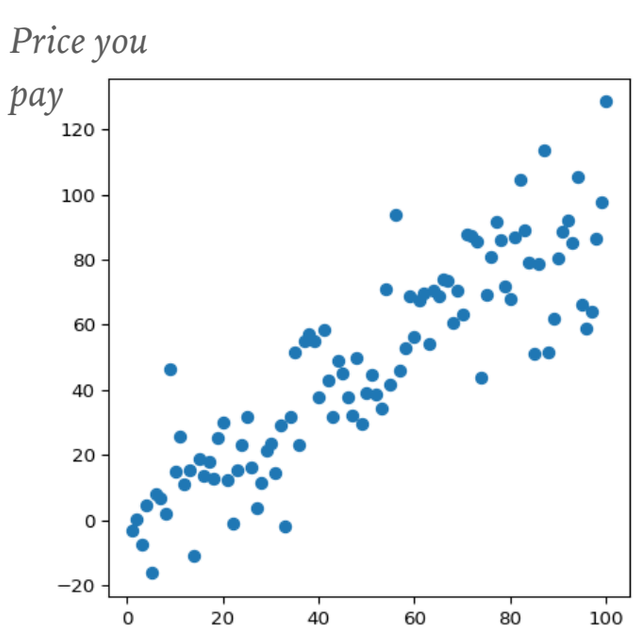


Of course it's not a precise model. A model is always an abstraction of the world.

x	*	done something with it	=	y
# of items	*	*some number*	=	*price*

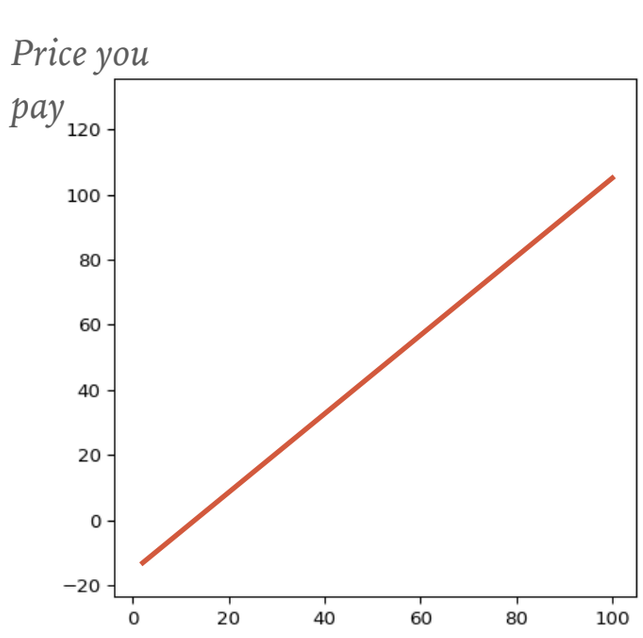
WHAT'S A LINEAR MODEL?

- *number of items* * m + b = *Price*
- predict the continuous target variable with another one
  - $mx + b = y$
  - -> explain the *price* by the *amount* you bought



Data

Fit a model

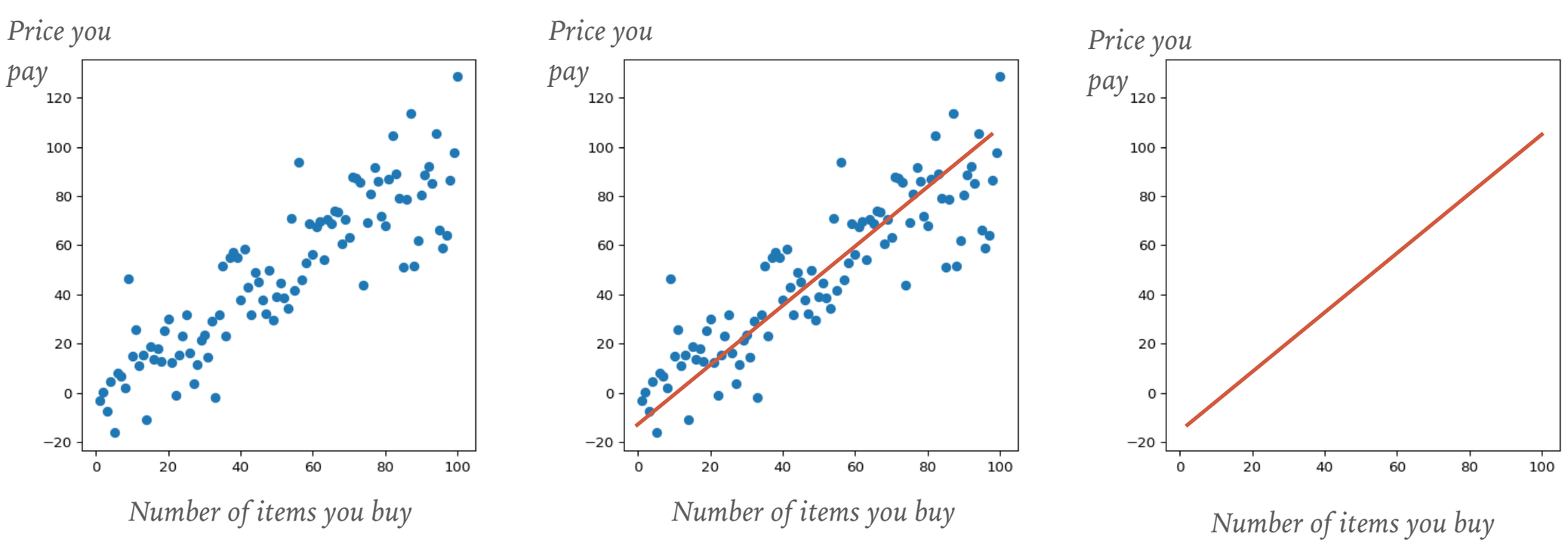


Model

WHAT'S A LINEAR MODEL?

- *number of items* * m + b = *Price*
- predict the continuous target variable with another one
  - $mx + b = y$

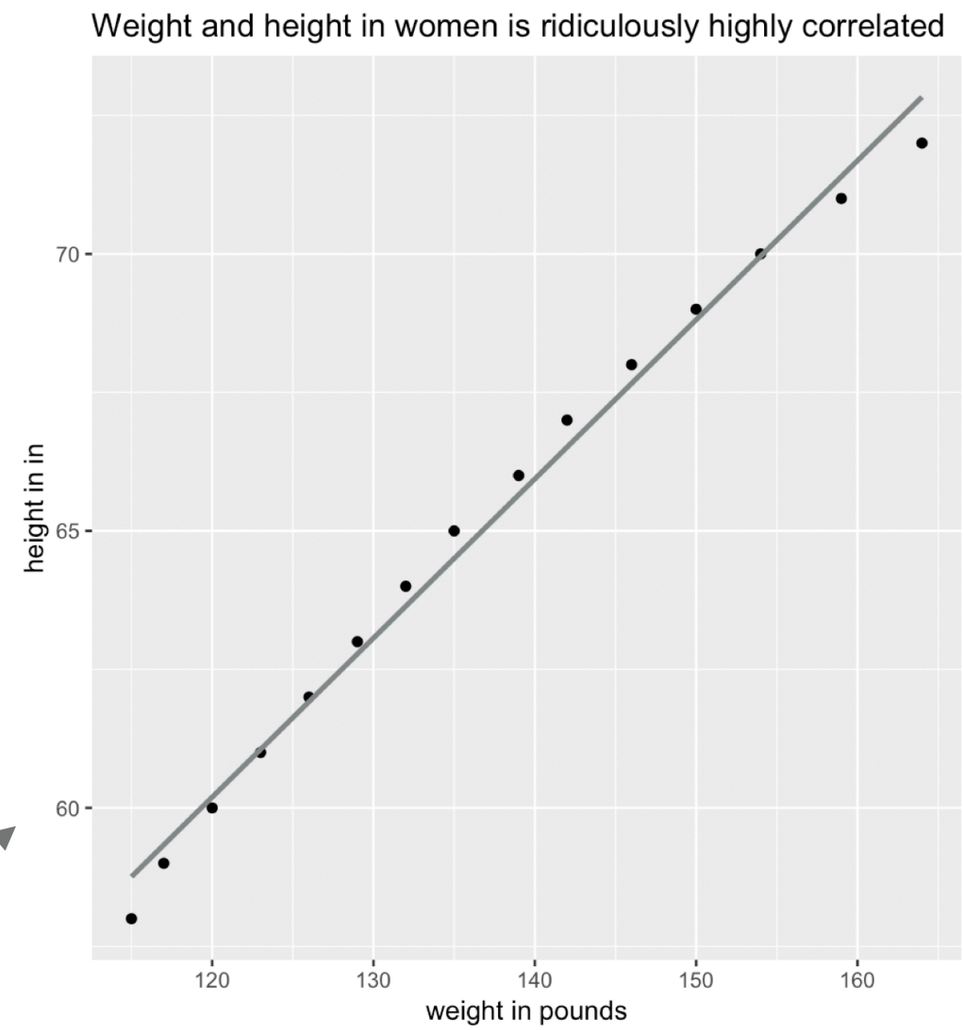
items -> * done something with it *ice* = by price *nount* you bought



Data → *Fit a model* → *Model*

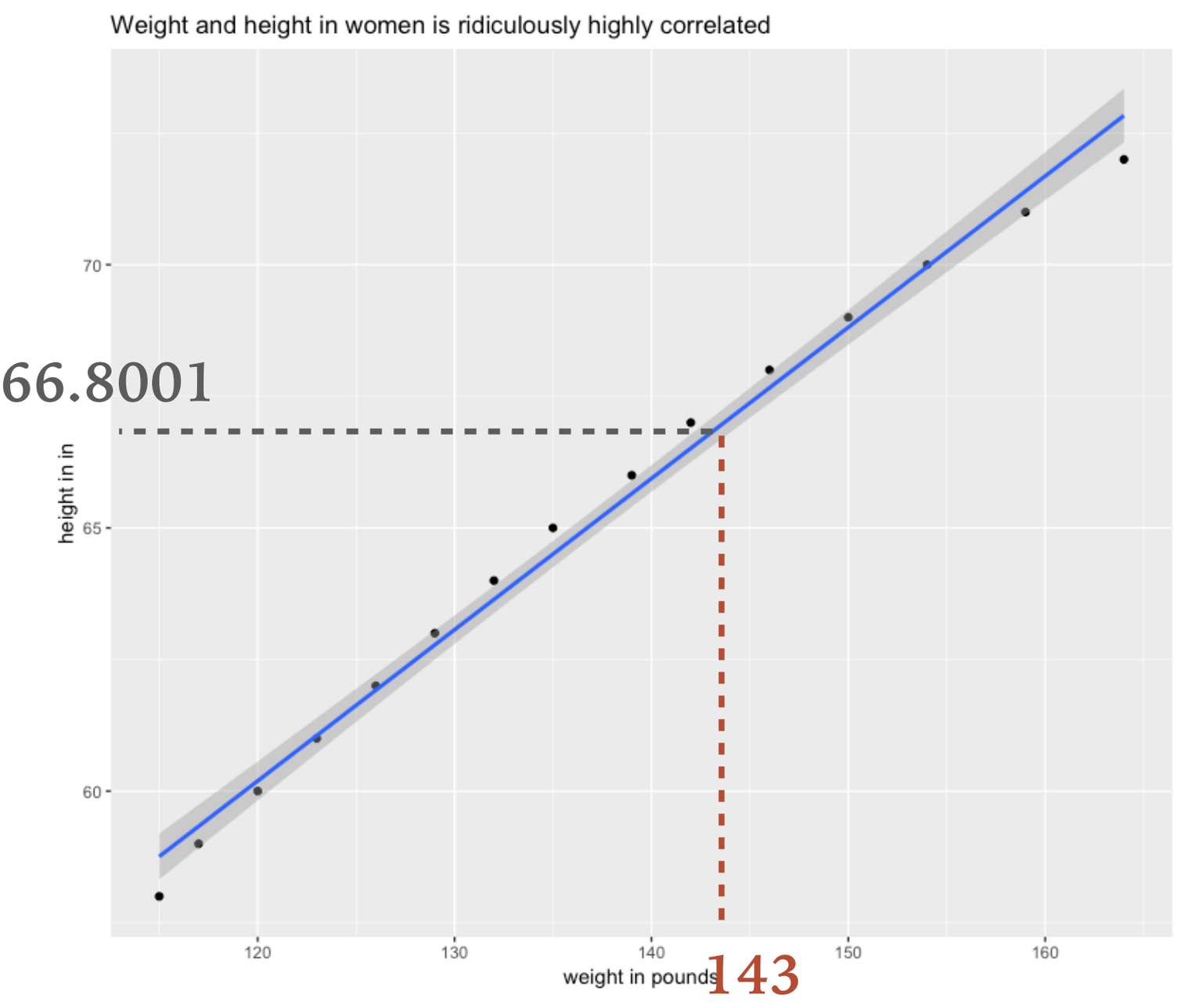
LET'S LOOK AT DATA(WOMEN).

	height	weight
1	58	115
2	59	117
3	60	120
4	61	123
5	62	126
6	63	129
7	64	132
8	65	135
9	66	139
10	67	142



What's the line?

THE LINEAR MODEL



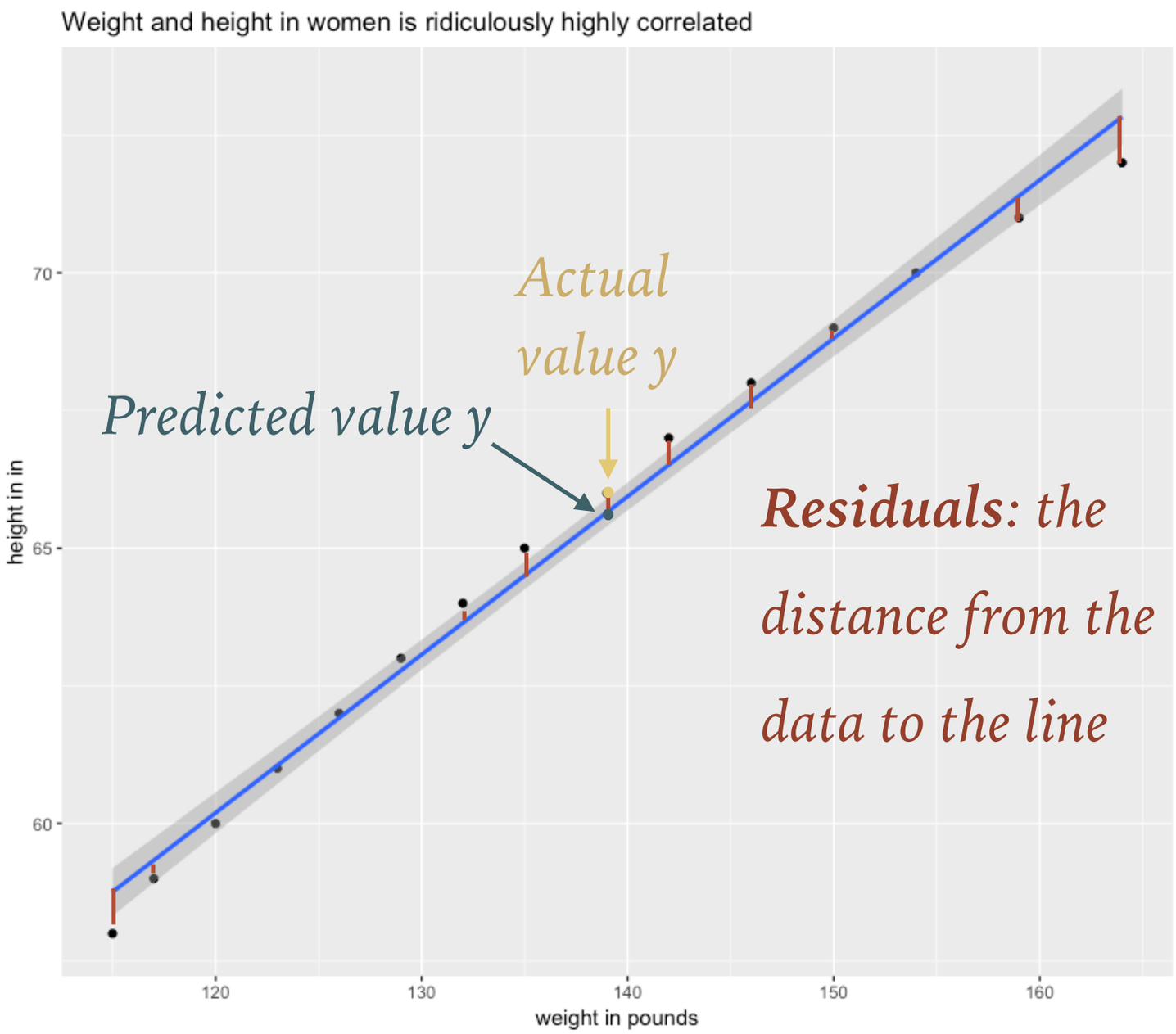
The **linear model** is the line that describes the data the best.

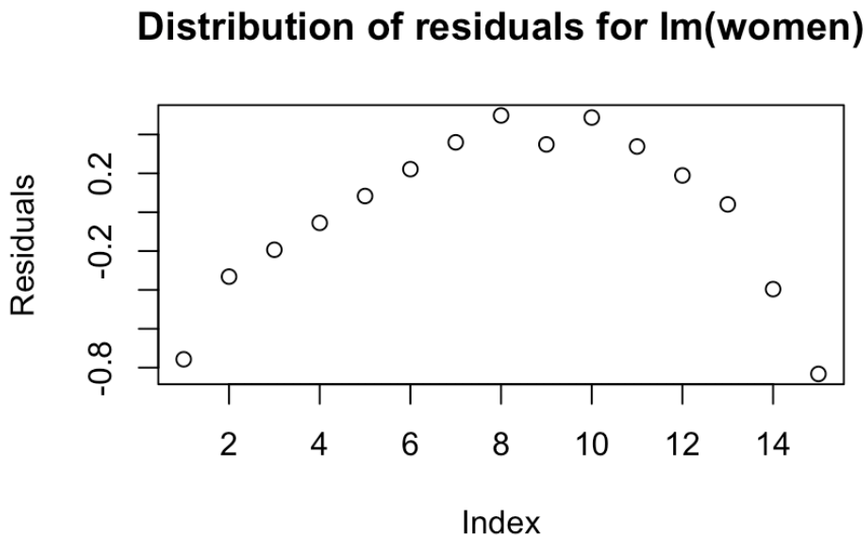
Let's say I am 143lbs (65kg).

Then, according to my model, I am 66.8001 in (170cm) tall.

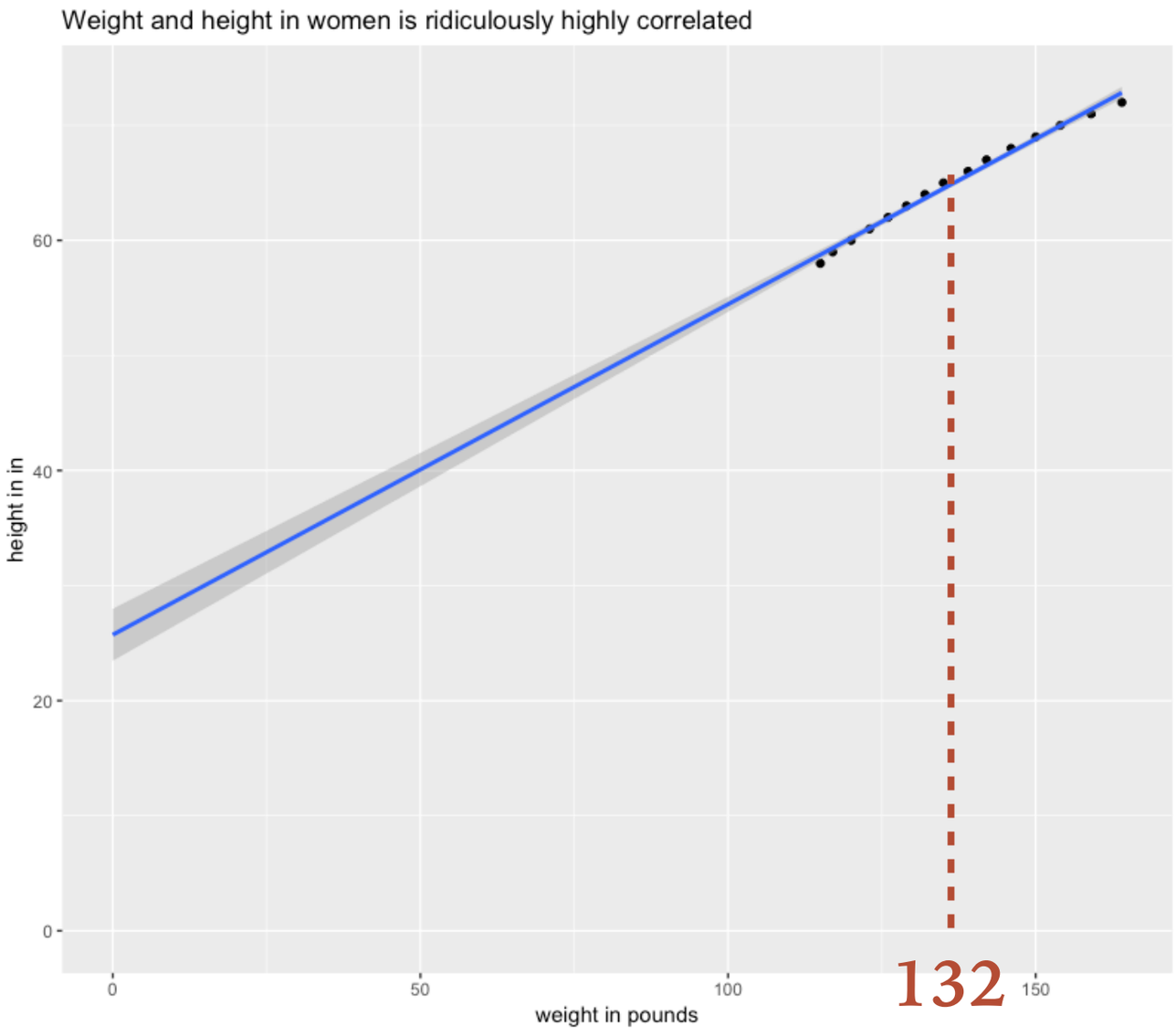
Anyone else?

THE RESIDUALS OF THE LINEAR MODEL





THE LINEAR MODEL PREDICTS Y VALUES.

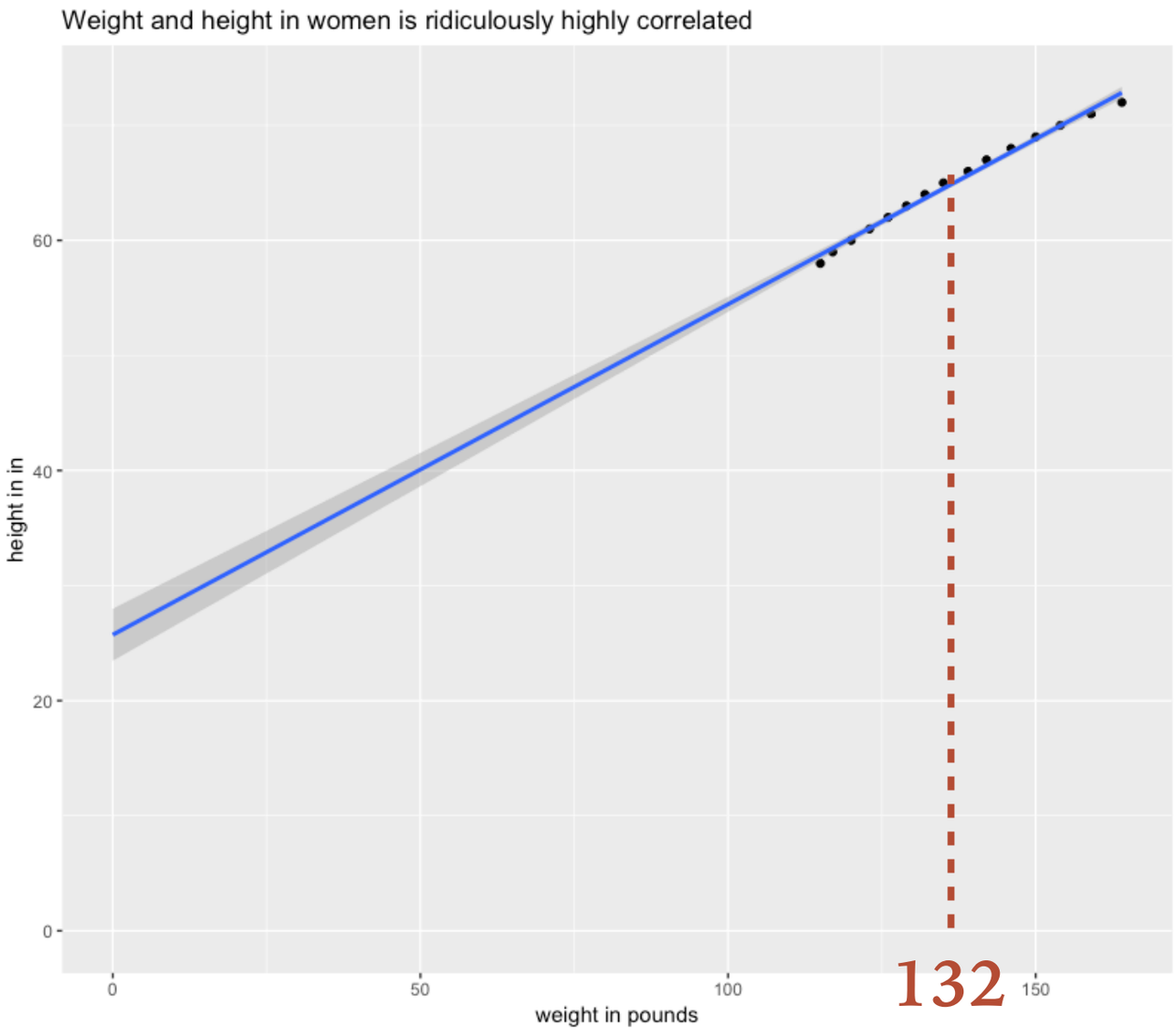


- Anna weighs approx. 132lbs (60kg).
- $mx + b = y$

Scaled to h
at (0, 0)

```
Call:
lm(formula = height ~ weight, data = women)

Coefficients:
(Intercept)       weight
    25.7235       0.2872
```

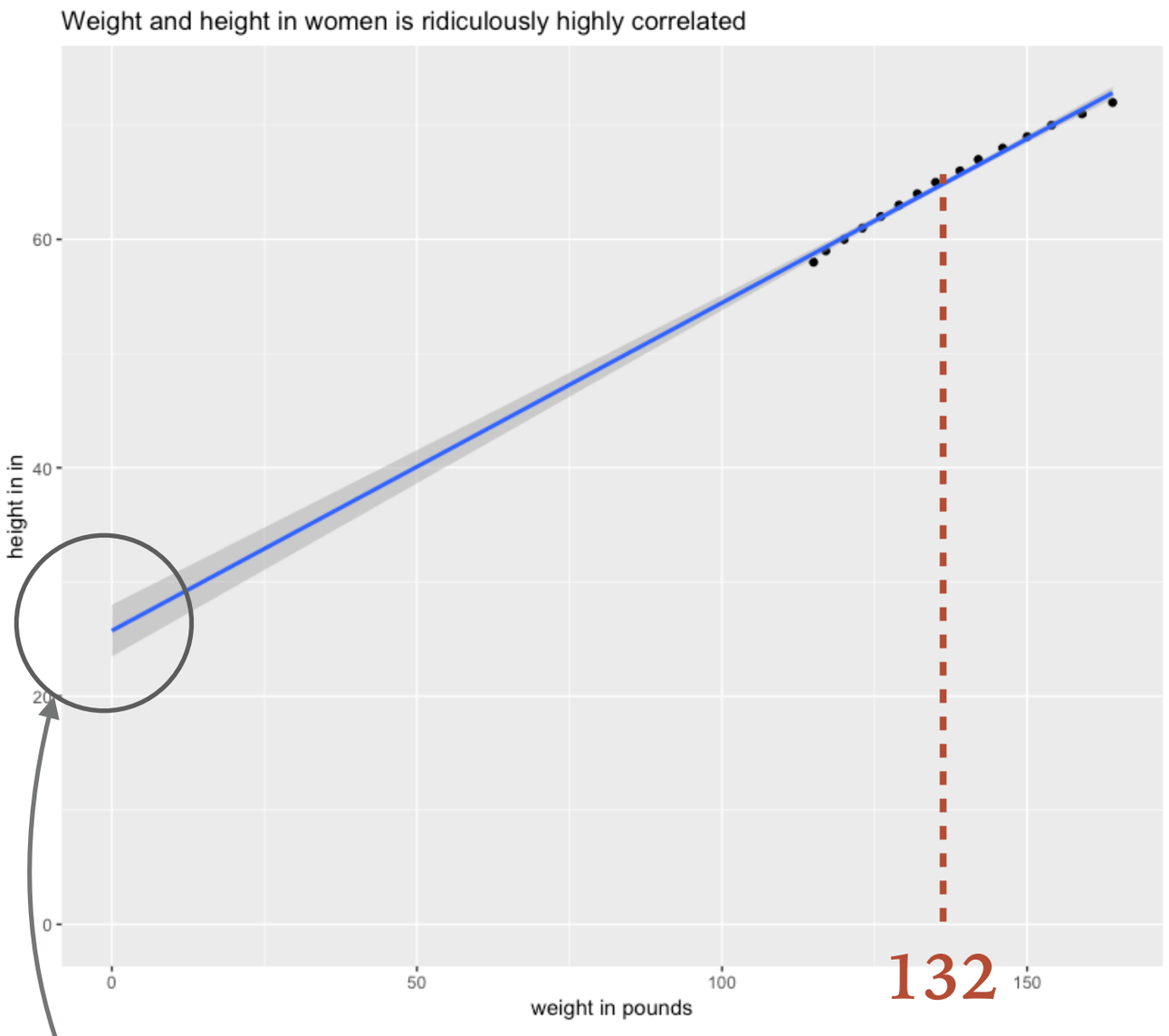
THE LINEAR MODEL PREDICTS Y VALUES.



- Anna weighs approx. 132lbs (60kg).
- $mx + b = y$
- 0.2872

```
Call:
lm(formula = height ~ weight, data = women)

Coefficients:
(Intercept)       weight
    25.7235       0.2872
```

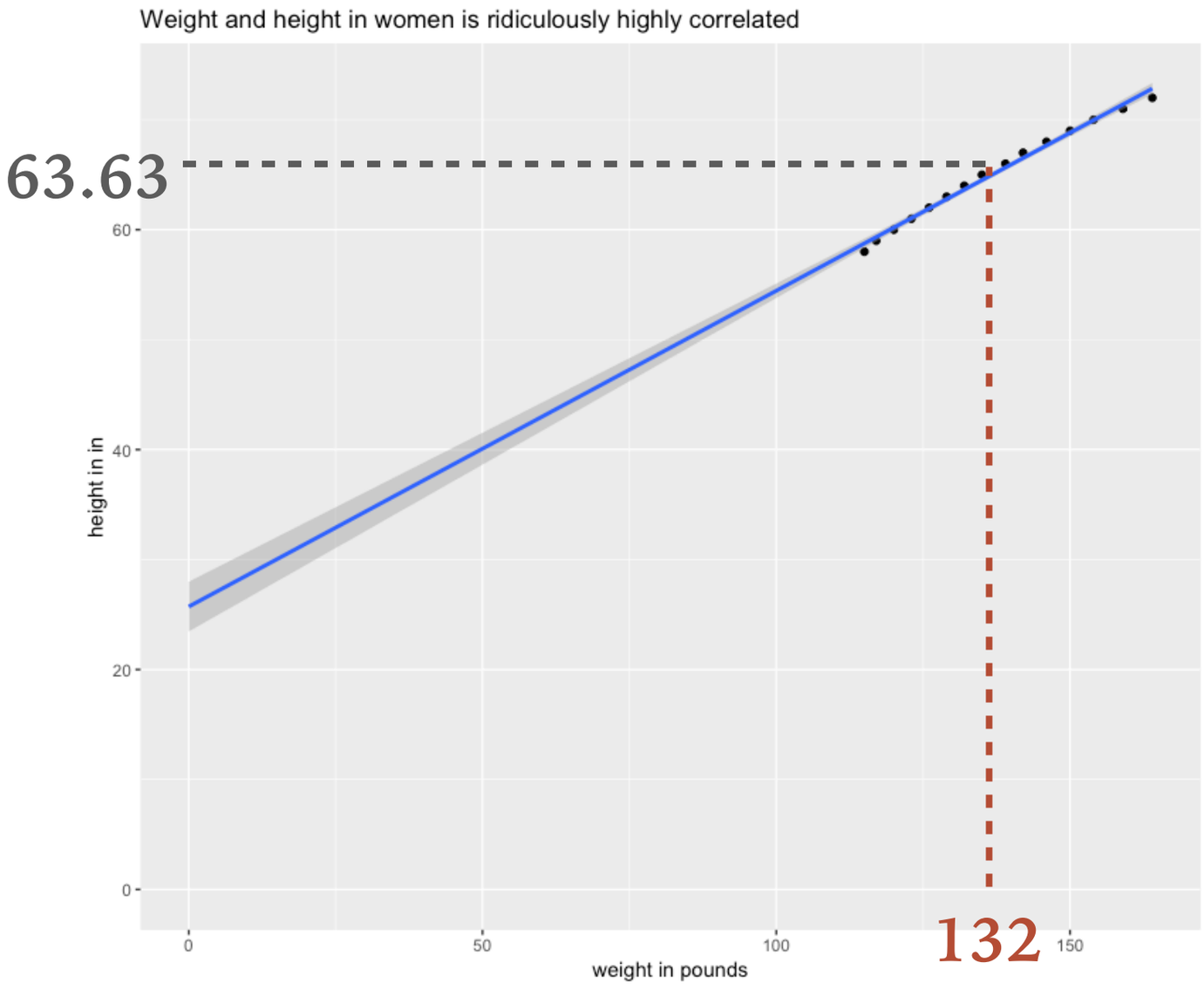
THE LINEAR MODEL PREDICTS Y VALUES.



- Anna weighs approx. 132lbs (60kg).
- $mx + b = y$
- 0.2872 * x + 25.723

```
Call:
lm(formula = height ~ weight, data = women)

Coefficients:
(Intercept)      weight
    25.7235      0.2872
```

THE LINEAR MODEL PREDICTS Y VALUES.



- Anna weighs approx. 132lbs (60kg).
- $mx + b = y$
- $0.2872 * 132 + 25.723 = 63.63$
- I must be 63.63 in (1,62m).

```
Call:
lm(formula = height ~ weight, data = women)

Coefficients:
(Intercept)      weight
    25.7235      0.2872
```

LET'S LOOK AT THE R OUTPUT

	height	weight
1	58	115
2	59	117
3	60	120
4	61	123
5	62	126
6	63	129
7	64	132
8	65	135
9	66	139
10	67	142

target variable

explained by weight

```
> lm(height ~ weight, data = women)

Call:
lm(formula = height ~ weight, data = women)

Coefficients:
(Intercept)       weight
    25.7235       0.2872
```

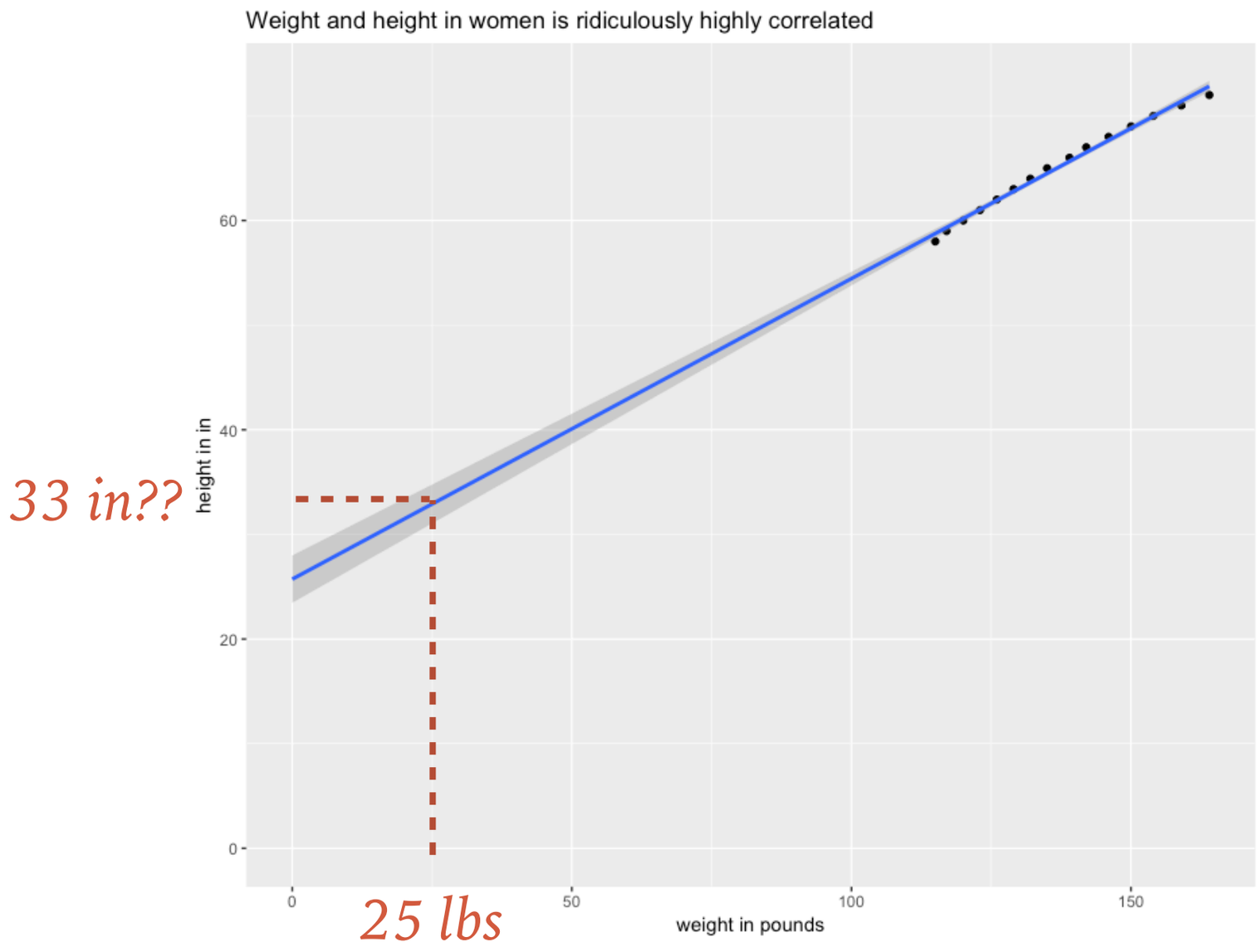
$$mx + b = y$$

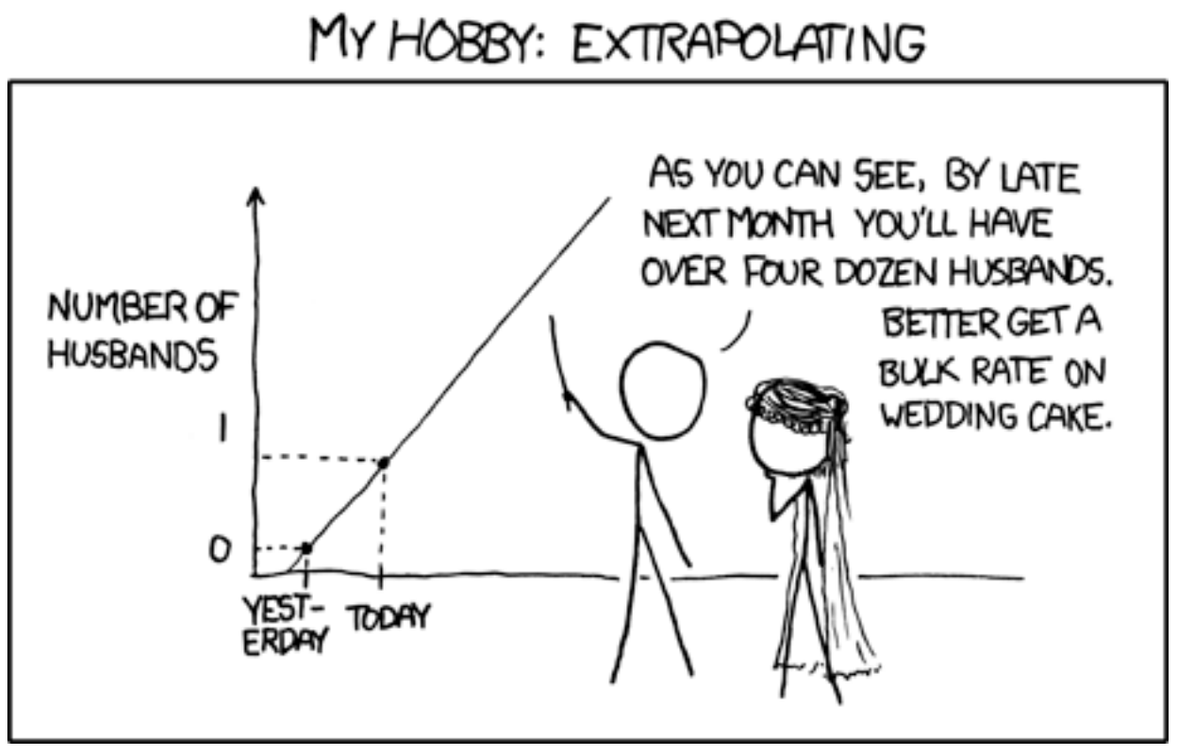
$$something * weight + something = height$$

PRACTICE

LAST BUT NOT LEAST: A WORD OF CAUTION

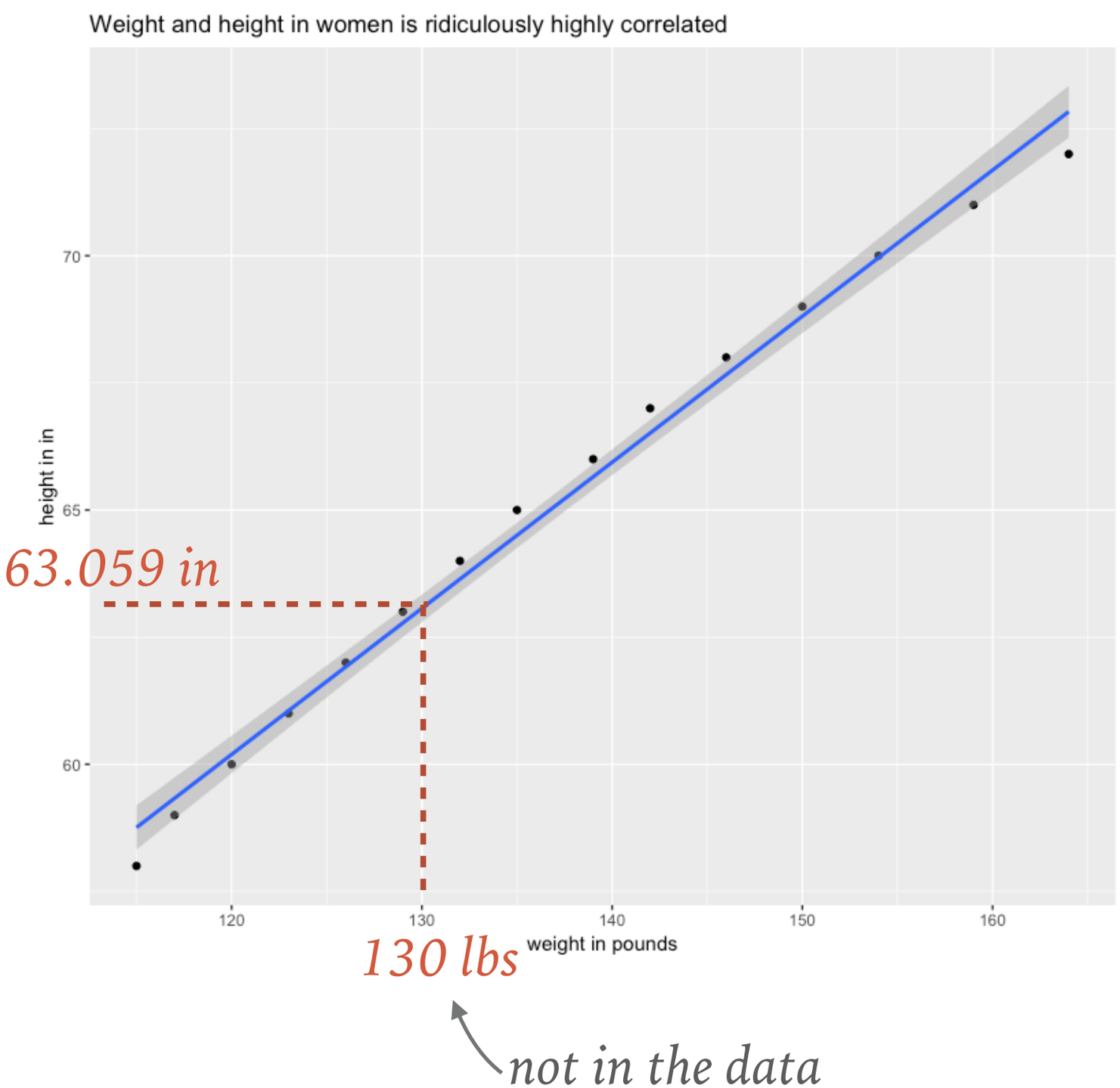
- Don't **extrapolate** out of the range of your data.





Source: xkcd.com/605

THOUGH THERE IS A THING CALLED INTERPOLATING.



- Though I don't have data for a person that is **130 lbs**, my model still gives me an estimate.

Next time: **Analysis of variance!**

See you there!

SOURCES

- grid by Flatart from the Noun Project
- banana by miracle from the Noun Project
- Apple by Lyhn from the Noun Project
- Bread by zidney from the Noun Project