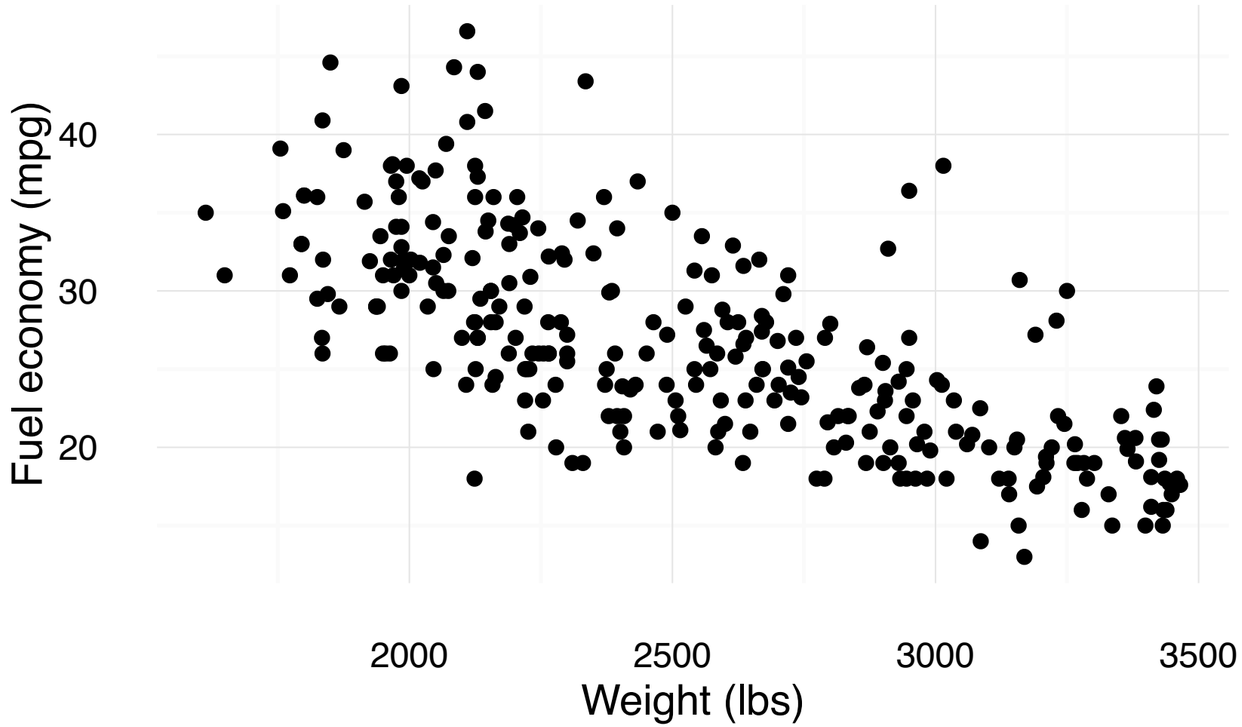
Correlation & Regression

Fuel economy data

Many factors go into determining what gas mileage a car will achieve

It's generally understood that heavier cars will get worse fuel economy, but it is not clear how much of an increase in weight will lead to a decrease in fuel economy

Scatterplots



Direction

Strength

Form/trend

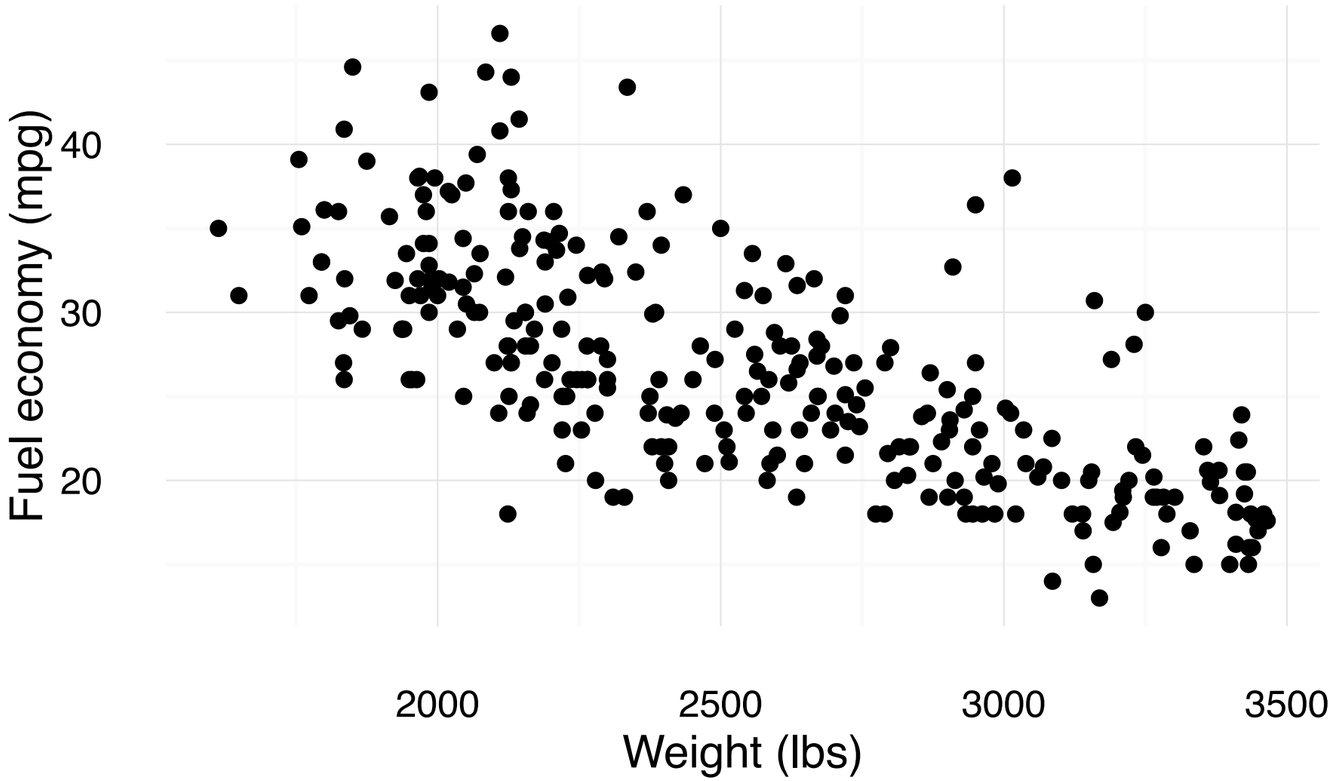
Unusual features

Correlation

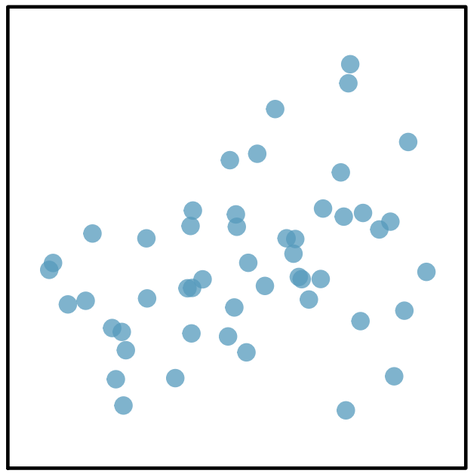
A numerical measure of the strength and direction of a linear association between two quantitative variables.

$$r = \sum \left(\frac{x - \overline{x}}{s_x} \right) \left(\frac{y - \overline{y}}{s_y} \right)$$

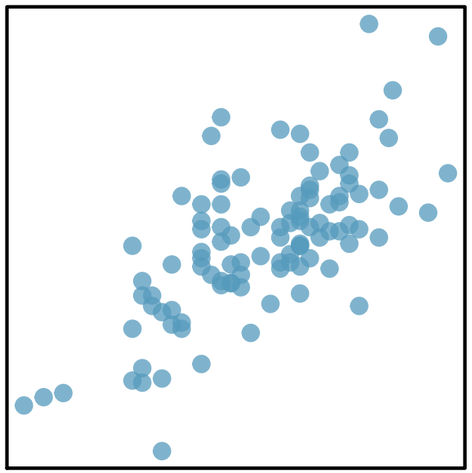
Using R, we find r = -0.71



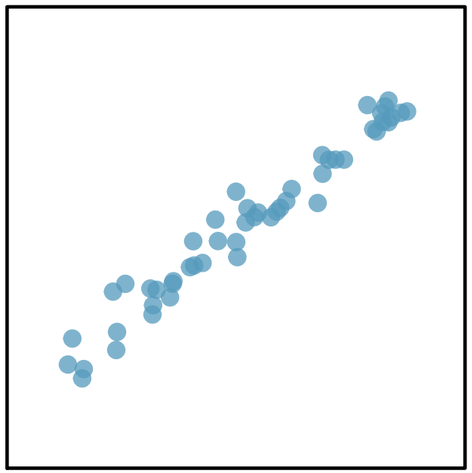

```
library(mosaic)
cor(MPG ~ Weight, data = mpg)
```



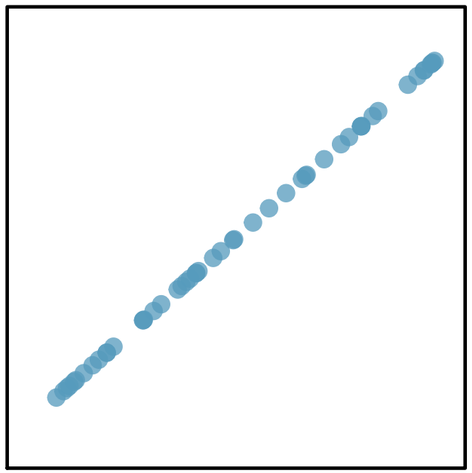
R = 0.33



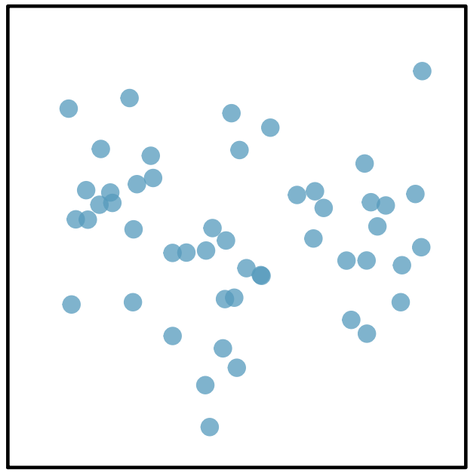
R = 0.69



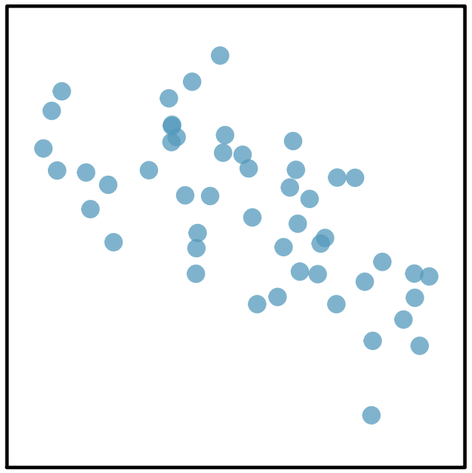
R = 0.98



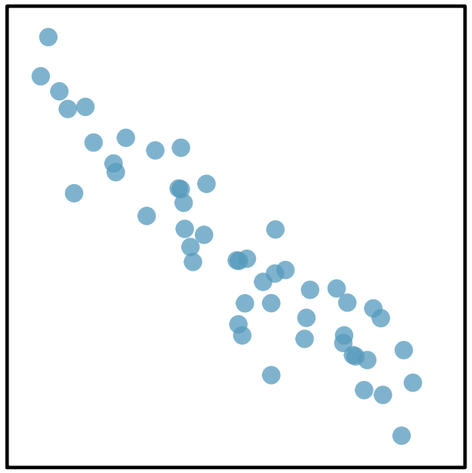
R = 1.00



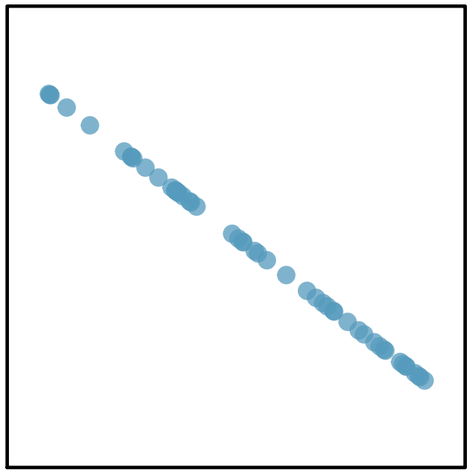
R = –0.08



R = –0.64

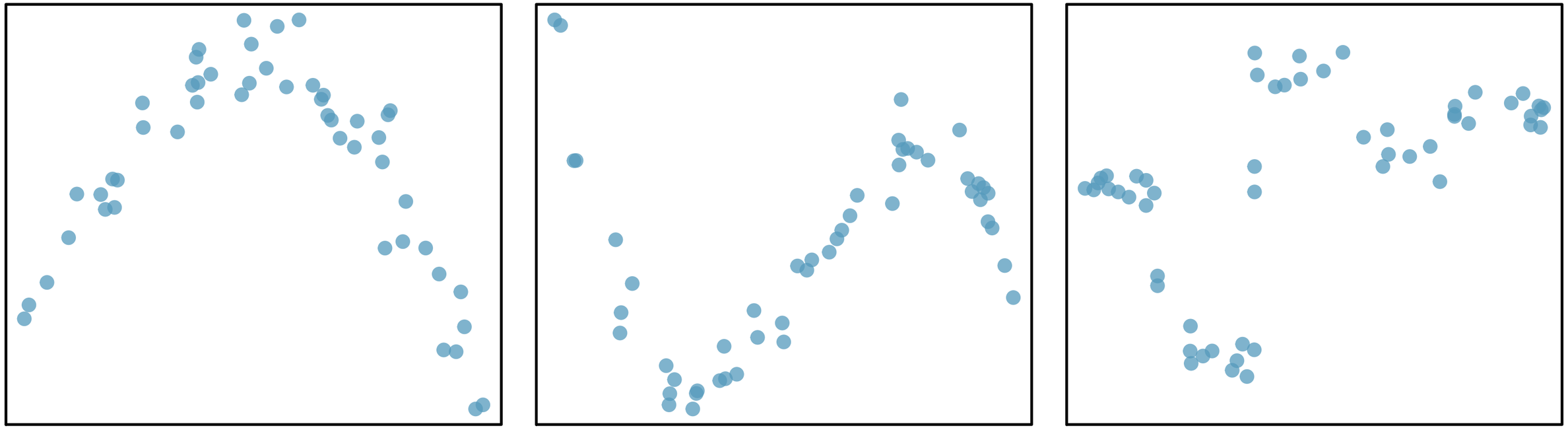


R = –0.92



R = –1.00

Beware of nonlinearity



R = –0.23

R = 0.31

R = 0.50

Properties

1. $-1 \leq r \leq 1$
2. Sign indicates direction
3. The closer r is to ± 1, the stronger the association
4. Unitless
5. Does not depend on the units of measurement
6. The correlation between x and y is the same as the correlation between y and x

Calibrating your intuition

mih5.github.io/statapps/correlation*game*/*correlation*game.html

guessthecorrelation.com

Cautions

1. Correlation can be heavily influenced by outliers. Don’t just look at the correlation! Always plot your data!
2. r = 0 indicates that there is no linear association between the two variables, but the variables could still be associated! Always plot your data!
3. Correlation does not imply causation! Remember to think!

Analytic goal

1. **Describe the relationship between weight and fuel economy**

2. **Predict fuel economy based on a vehicle's weight**

Fitted regression equation

$$\hat{y} = b_0 + b_1 x$$

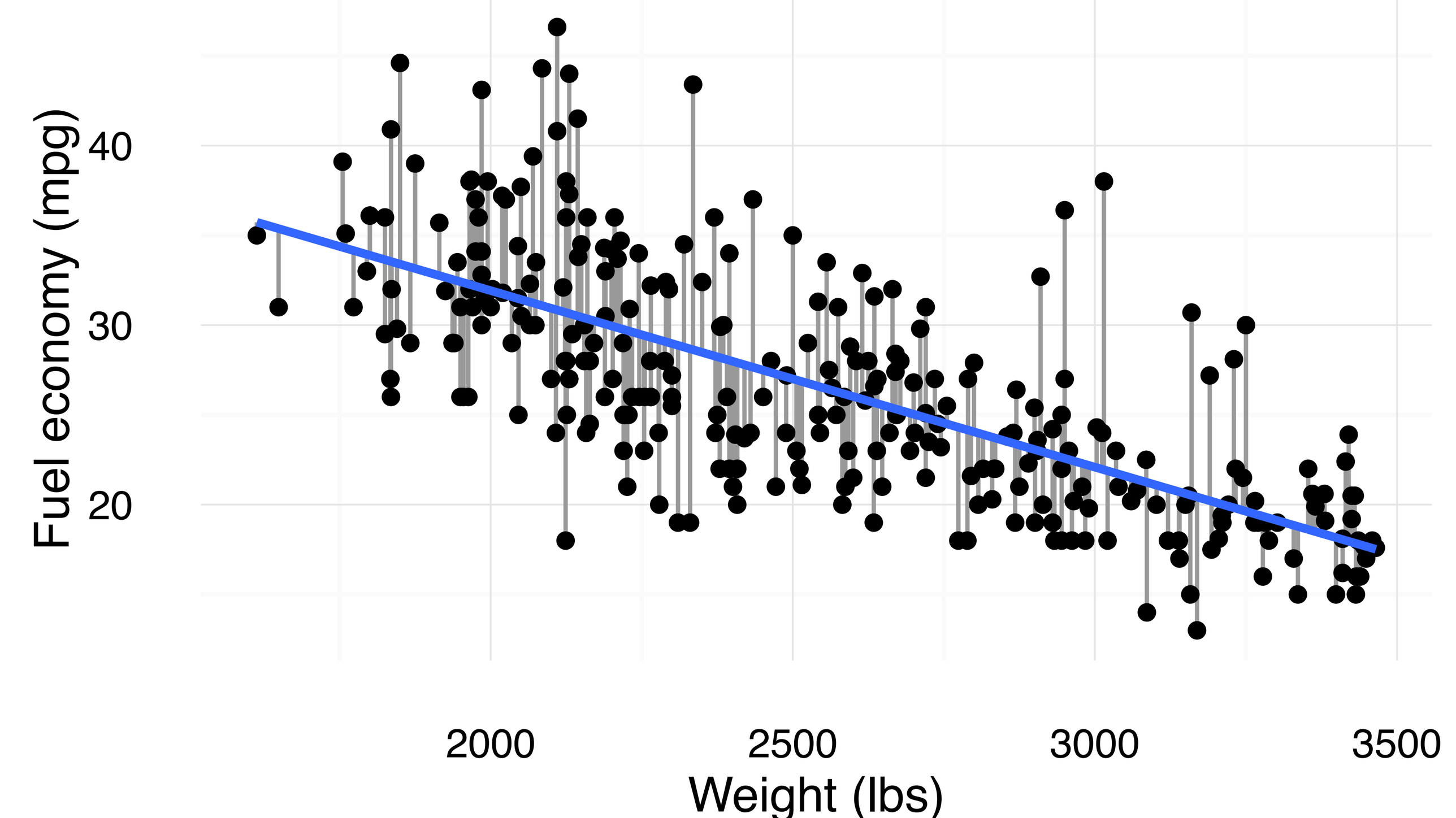
Least squares (LS) regression line

Residuals are the leftovers from the model fit:

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

The LS regression line minimizes the sum of the squared residuals:

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$


Fuel economy (mpg)
40
30
20
2000
2500
3000
3500
Weight (lbs)

```
mod <- lm(MPG ~ Weight, data = mpg)
summary(mod)
```

```
Call:
lm(formula = MPG ~ Weight, data = mpg)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.7011  -3.3404  -0.5987   2.3588  16.0605 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 51.5871689  1.4835394   34.77   <2e-16 ***
Weight      -0.0098334  0.0005749  -17.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.723 on 287 degrees of freedom
Multiple R-squared:  0.5048,    Adjusted R-squared:  0.5031 
F-statistic: 292.6 on 1 and 287 DF,  p-value: < 2.2e-16
```

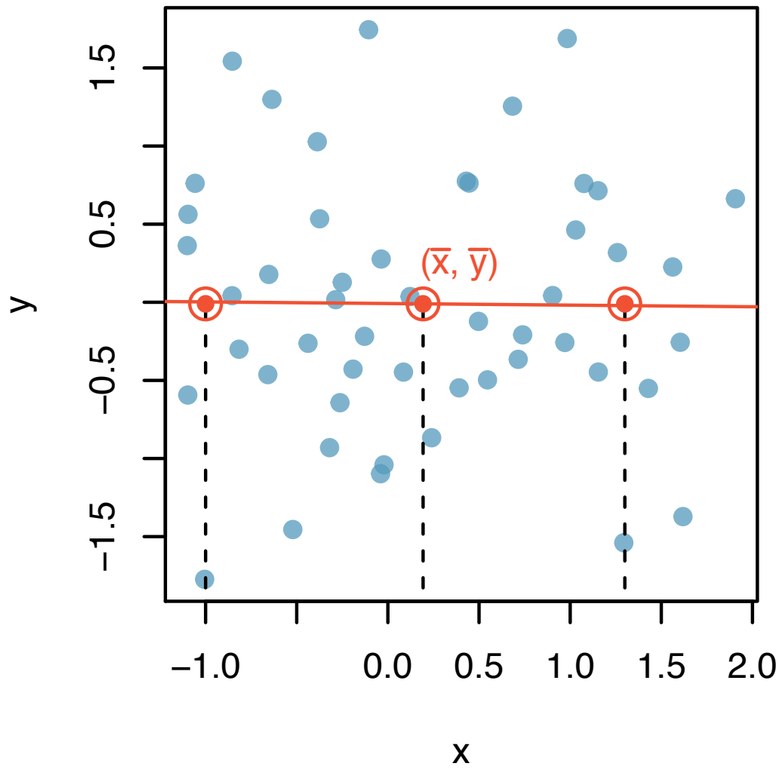
Interpreting the slope

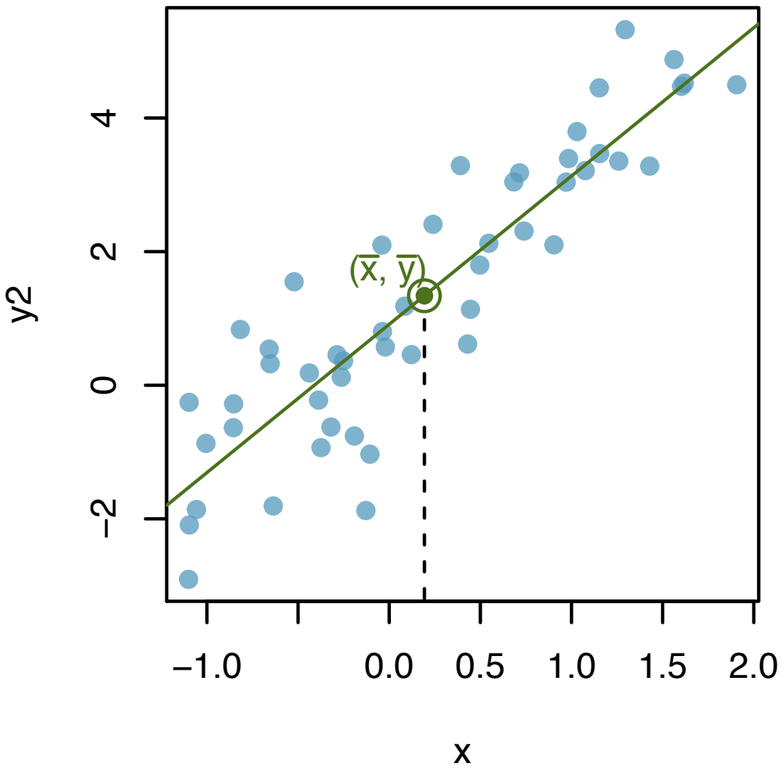
For each unit increase in x, y is expected to be higher/lower on average by the slope.

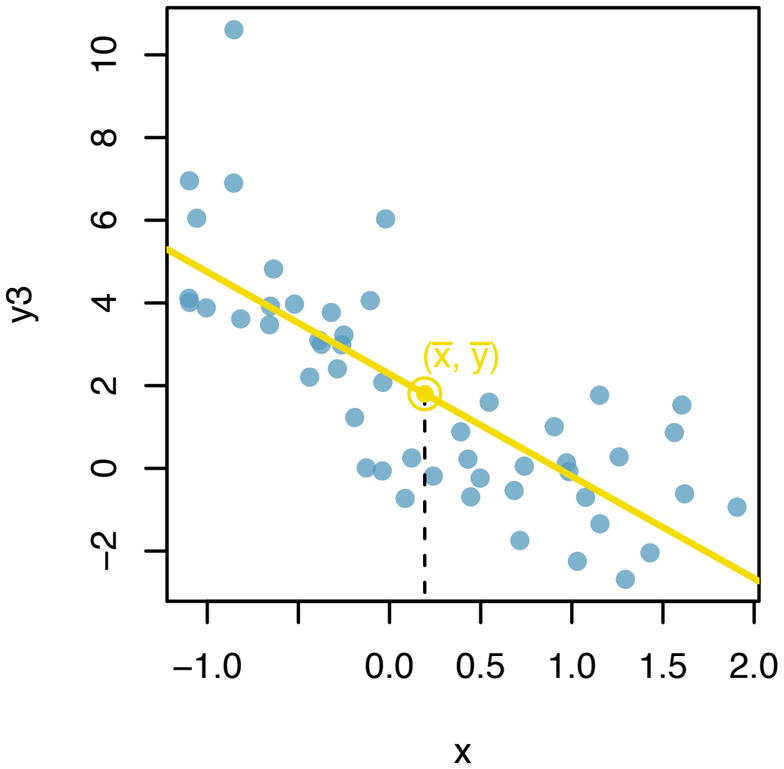
Interpreting the intercept

When x = 0, y is expected to equal the intercept.

The LS regresion line always passes through $(\overline{x}, \overline{y})$







Predict

How would we use the model to predict the fuel economy for a car weighing 2,500 lbs?

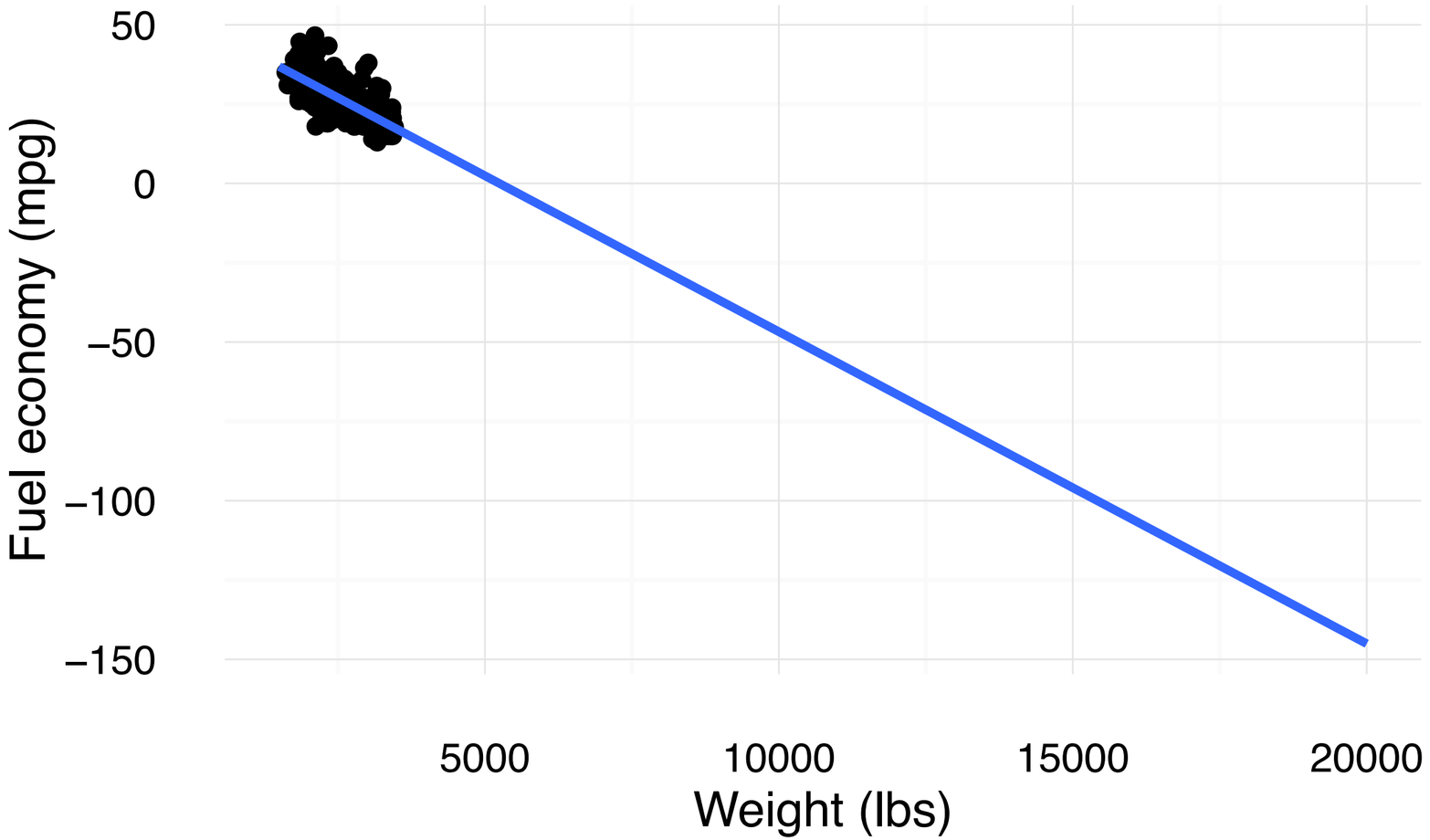
How do we interpret this number?

Predict

How would we use the model to predict the fuel economy for a semi weighing 20,000 lbs?

Does this prediction make sense?

Don't extrapolate



Sometimes the intercept might be an extrapolation: useful for adjusting the height of the line, but meaningless in the context of the data.