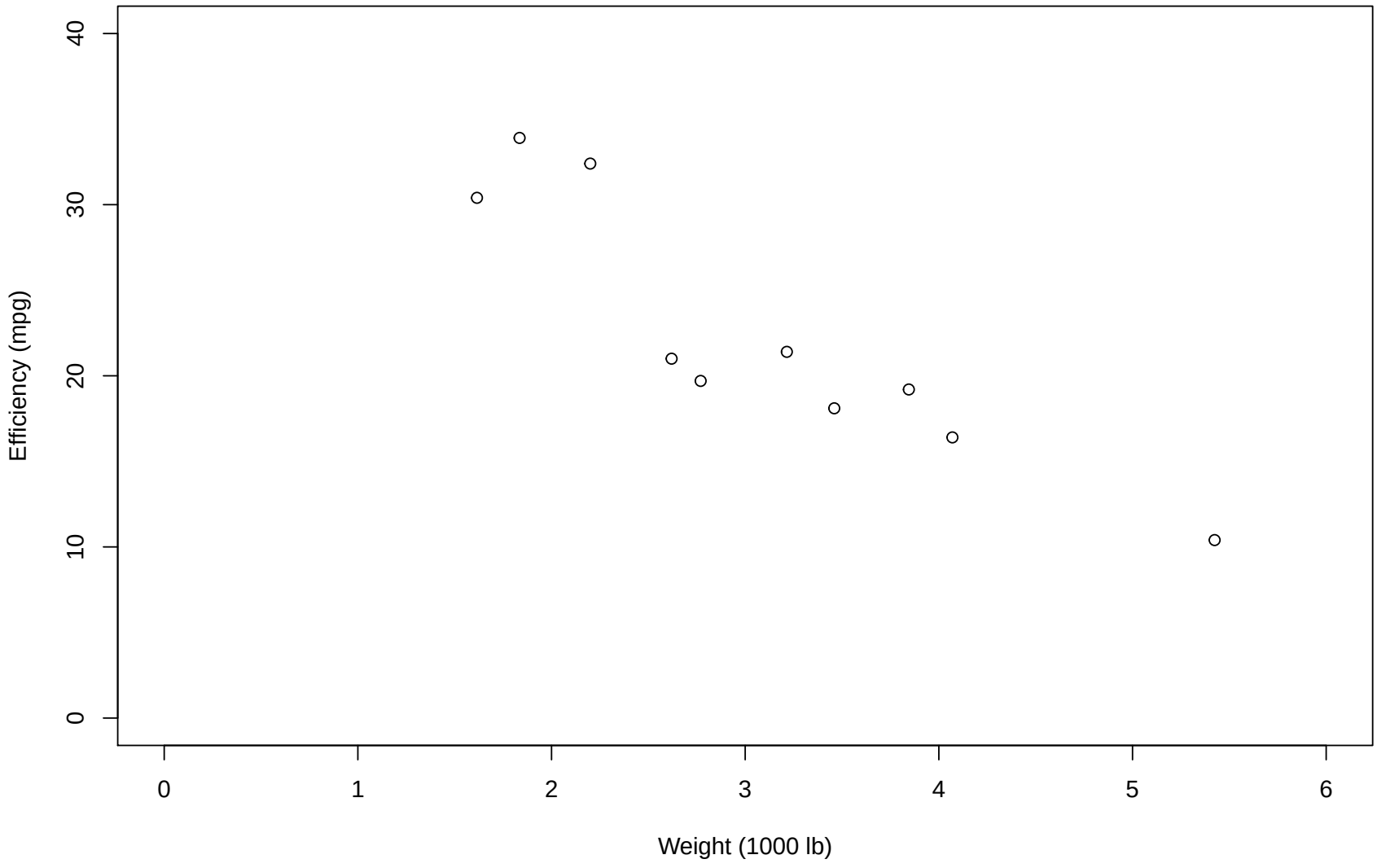
# Two (possibly dependent) variables

Consider the following data from 1974 Motor Trend US magazine:

| weight (1000s lb) | efficiency (mpg) |
|---|---|
| 3.215 | 21.4 |
| 1.835 | 33.9 |
| 1.615 | 30.4 |
| 2.77 | 19.7 |
| 3.845 | 19.2 |
| 2.2 | 32.4 |
| 2.62 | 21 |
| 3.46 | 18.1 |
| 5.424 | 10.4 |
| 4.07 | 16.4 |

# Make a scatterplot



# Find the (Pearson's) correlation coefficient

The correlation coefficient has many equivalent formulas.

The correlation coefficient can be thought of as the mean of products of standard scores.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

The following is probably easier to use for calculations.

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

# Calculate the correlation coefficient

| $x$ | $y$ | $xy$ |
|---|---|---|
| 3.215 | 21.4 | 68.801 |
| 1.835 | 33.9 | 62.2065 |
| 1.615 | 30.4 | 49.096 |
| 2.77 | 19.7 | 54.569 |
| 3.845 | 19.2 | 73.824 |
| 2.2 | 32.4 | 71.28 |
| 2.62 | 21 | 55.02 |
| 3.46 | 18.1 | 62.626 |
| 5.424 | 10.4 | 56.4096 |
| 4.07 | 16.4 | 66.748 |
| $\sum x = 31.054$ | $\sum y = 222.9$ | $\sum x_i y_i = 620.5801$ |
| $\bar{x} = 3.1054$ | $\bar{y} = 22.29$ | |
| $s_x = 1.15$ | $s_y = 7.57$ | |

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{620.5801 - 10 \cdot 3.1054 \cdot 22.29}{(10-1) \cdot 1.15 \cdot 7.57} = -0.911$$

# Determine the best-fit line

The regression line has the form

$$y = a + bx$$

So, $a$ is the $y$-intercept and $b$ is the slope. We have formulas to determine them:

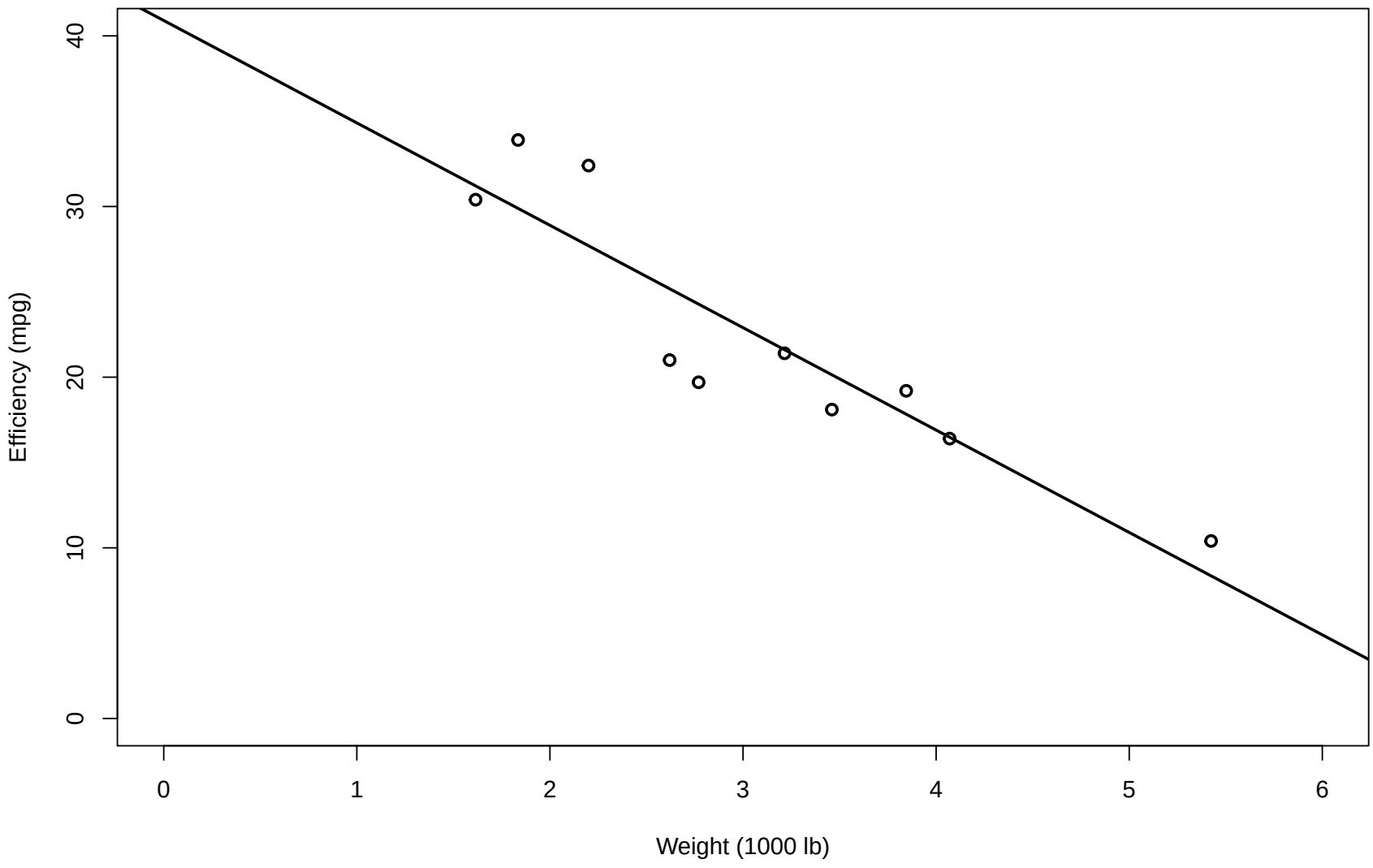
$$b = r\frac{s_y}{s_x} = -0.911 \cdot \frac{7.57}{1.15} = -6$$

$$a = \bar{y} - b\bar{x} = 22.29 - (-6) \cdot 3.1054 = 40.9$$

Our regression line:

$$y = 40.9 + (-6)x$$

# Graph the regression line



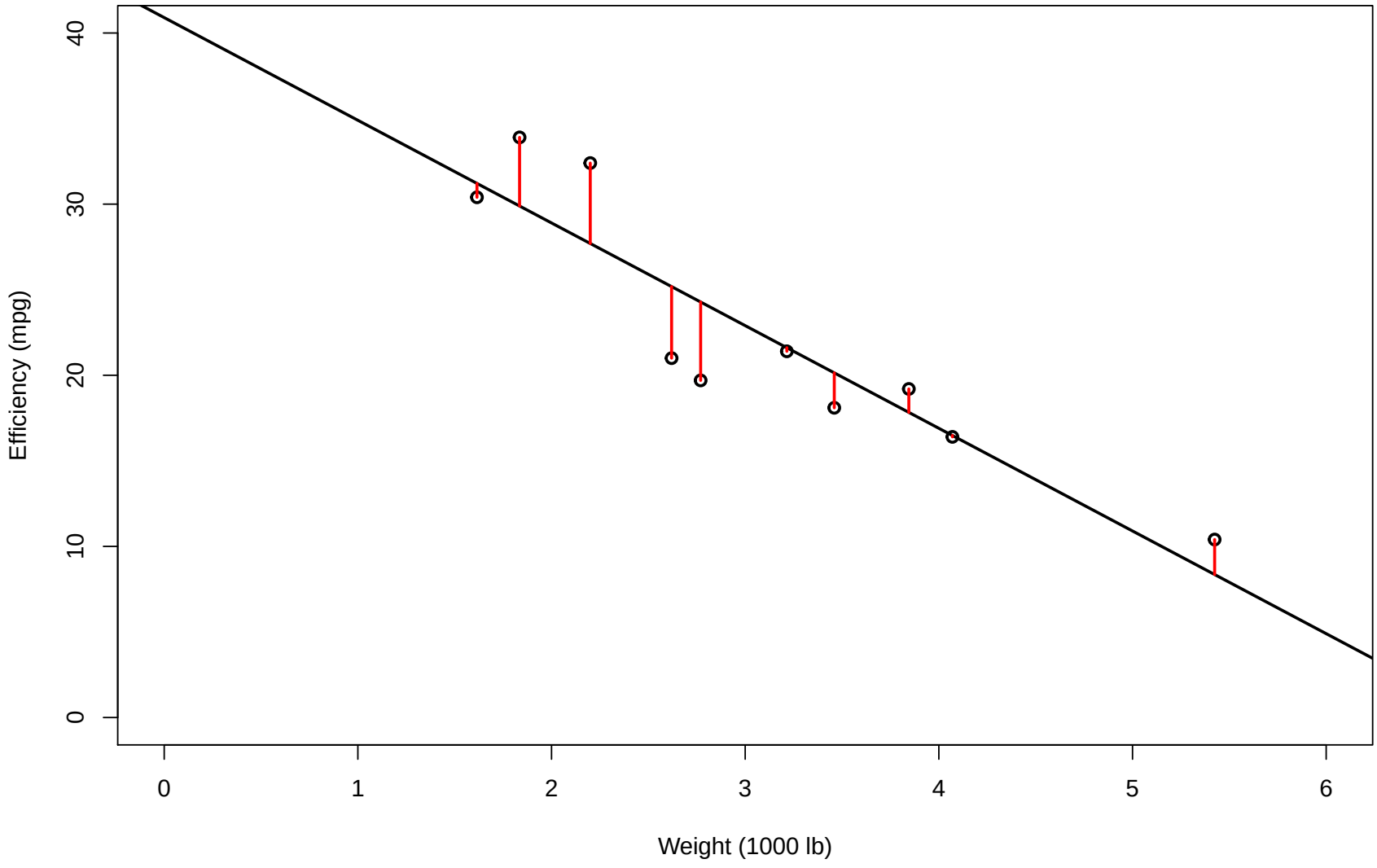
# Calculate and interpret the coefficient of determination

The coefficient of determination is the square of the correlation coefficient.

$$\text{coefficient of determination} = r^2 = (-0.911)^2 = 0.83$$

We say that 83% of the variance in $y$ (fuel efficiency) is explained by $x$ (weight).

# Determine the residuals (differences between actual and predicted)



# Determine the residuals and standard error of regression

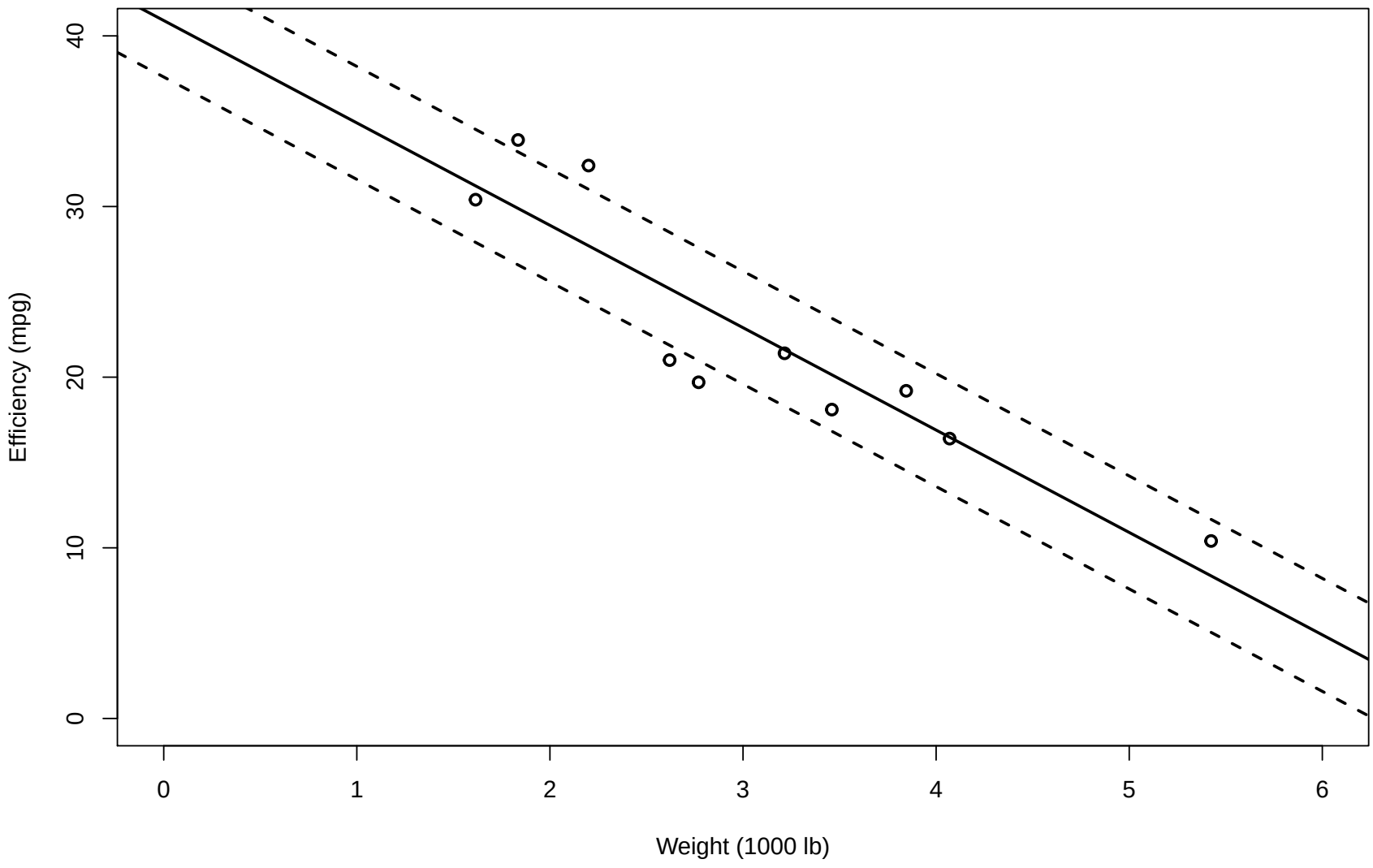
Let $y'$ be the predicted value from $y = a + bx$ linear model.

| $x$ | $y$ | $y'$ | $y - y'$ | $(y - y')^2$ |
|---|---|---|---|---|
| 3.215 | 21.4 | 21.6 | -0.2 | 0.04 |
| 1.835 | 33.9 | 29.9 | 4 | 16 |
| 1.615 | 30.4 | 31.2 | -0.8 | 0.64 |
| 2.77 | 19.7 | 24.3 | -4.6 | 21.2 |
| 3.845 | 19.2 | 17.8 | 1.4 | 1.96 |
| 2.2 | 32.4 | 27.7 | 4.7 | 22.1 |
| 2.62 | 21 | 25.2 | -4.2 | 17.6 |
| 3.46 | 18.1 | 20.1 | -2 | 4 |
| 5.424 | 10.4 | 8.36 | 2.04 | 4.16 |
| 4.07 | 16.4 | 16.5 | -0.1 | 0.01 |

$$s_{y-y'} = \sqrt{\frac{\sum(y - y')^2}{n - 2}} = \sqrt{\frac{87.71}{10 - 2}} = 3.31$$

We think predictions tend to be off by 3.31 mpg.

# Draw the ± standard error of regression



# Make a point estimate using the linear model

Predict the fuel efficiency of a car with a mass of 200 lbs ($x = 0.2$).

$$y = 40.9 + (-6)(0.2) = 39.7$$

. . . notice this model does not do a good job predicting the fuel efficiency of a bicycle. A bicycle is much more efficient that that (approximately 700 mpg on an energy basis).