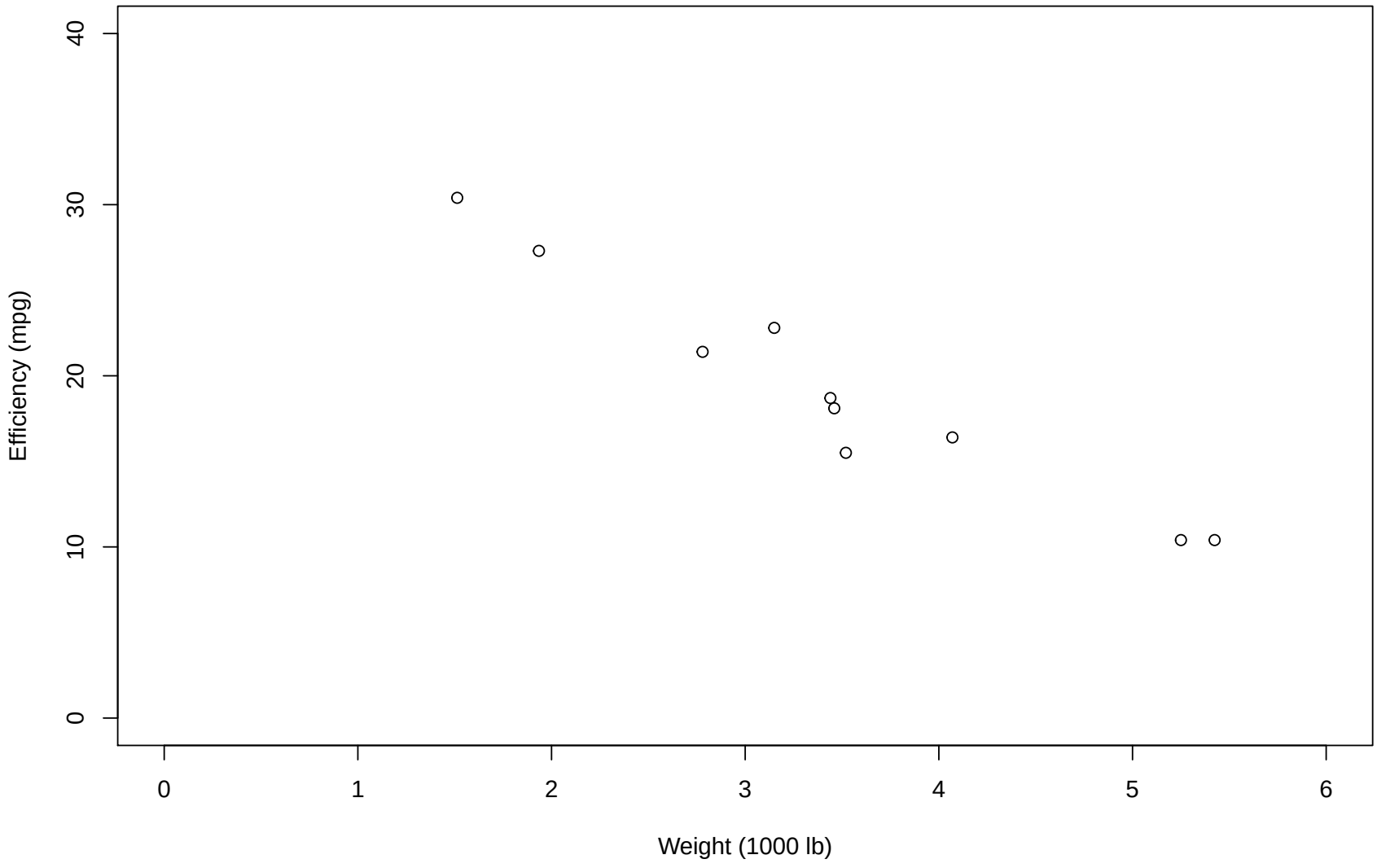
# Two (possibly dependent) variables

Consider the following data from 1974 Motor Trend US magazine:

| weight (1000s lb) | efficiency (mpg) |
|---|---|
| 1.513 | 30.4 |
| 5.424 | 10.4 |
| 1.935 | 27.3 |
| 3.52 | 15.5 |
| 3.44 | 18.7 |
| 4.07 | 16.4 |
| 5.25 | 10.4 |
| 3.15 | 22.8 |
| 2.78 | 21.4 |
| 3.46 | 18.1 |

# 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

| $x$ | $y$ | $xy$ |
|---|---|---|
| 1.513 | 30.4 | 45.9952 |
| 5.424 | 10.4 | 56.4096 |
| 1.935 | 27.3 | 52.8255 |
| 3.52 | 15.5 | 54.56 |
| 3.44 | 18.7 | 64.328 |
| 4.07 | 16.4 | 66.748 |
| 5.25 | 10.4 | 54.6 |
| 3.15 | 22.8 | 71.82 |
| 2.78 | 21.4 | 59.492 |
| 3.46 | 18.1 | 62.626 |
| $\sum x = 34.542$ | $\sum y = 191.4$ | $\sum x_i y_i = 589.4043$ |
| $\bar{x} = 3.4542$ | $\bar{y} = 19.14$ | |
| $s_x = 1.25$ | $s_y = 6.55$ | |

$$r = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y} = \frac{589.4043 - 10 \cdot 3.4542 \cdot 19.14}{(10-1) \cdot 1.25 \cdot 6.55} = -0.971$$

# 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:

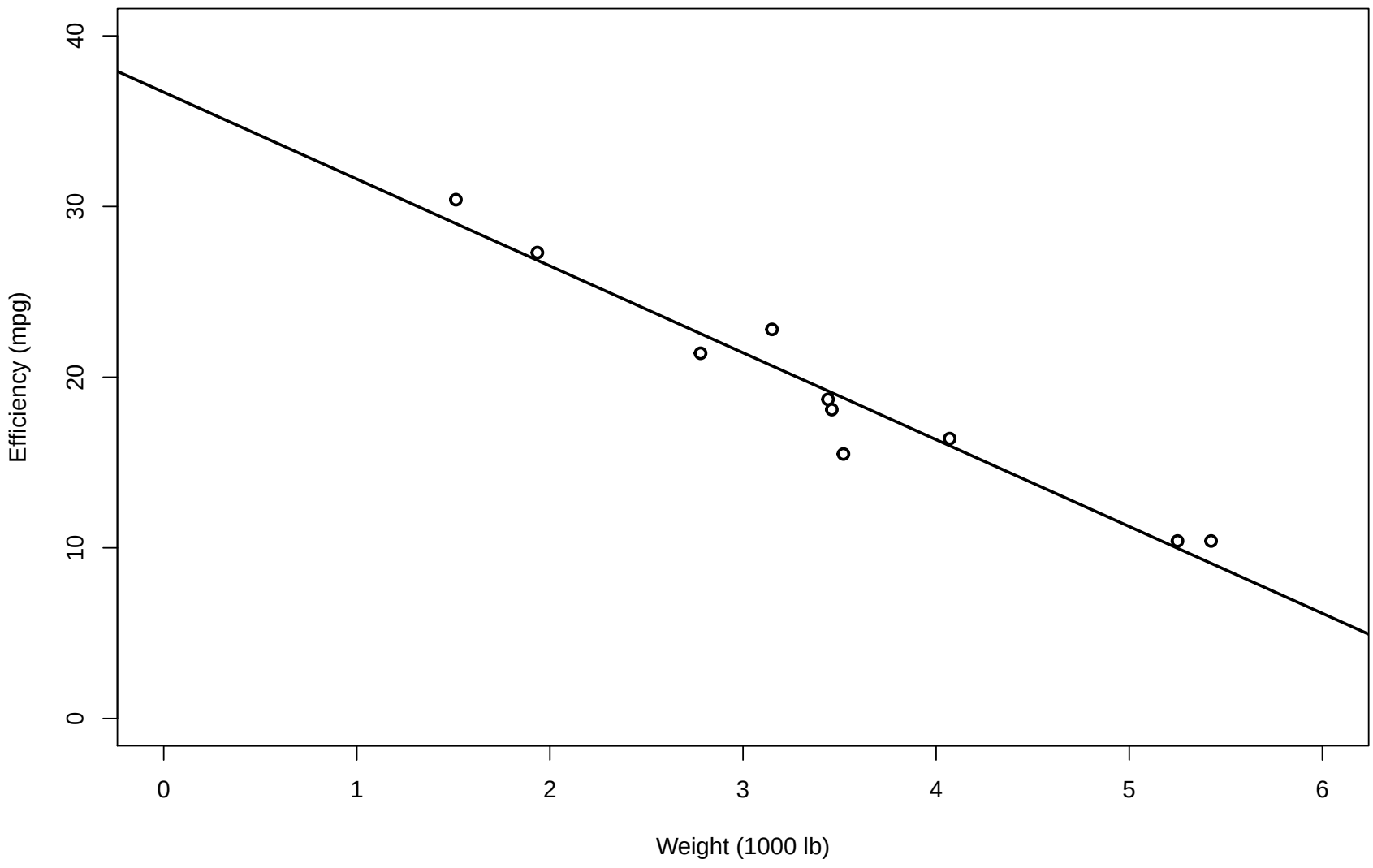
$$b = r\frac{s_y}{s_x} = -0.971 \cdot \frac{6.55}{1.25} = -5.09$$

$$a = \bar{y} - b\bar{x} = 19.14 - (-5.09) \cdot 3.4542 = 36.7$$

Our regression line:

$$y = 36.7 + (-5.09)x$$

# Graph the regression line



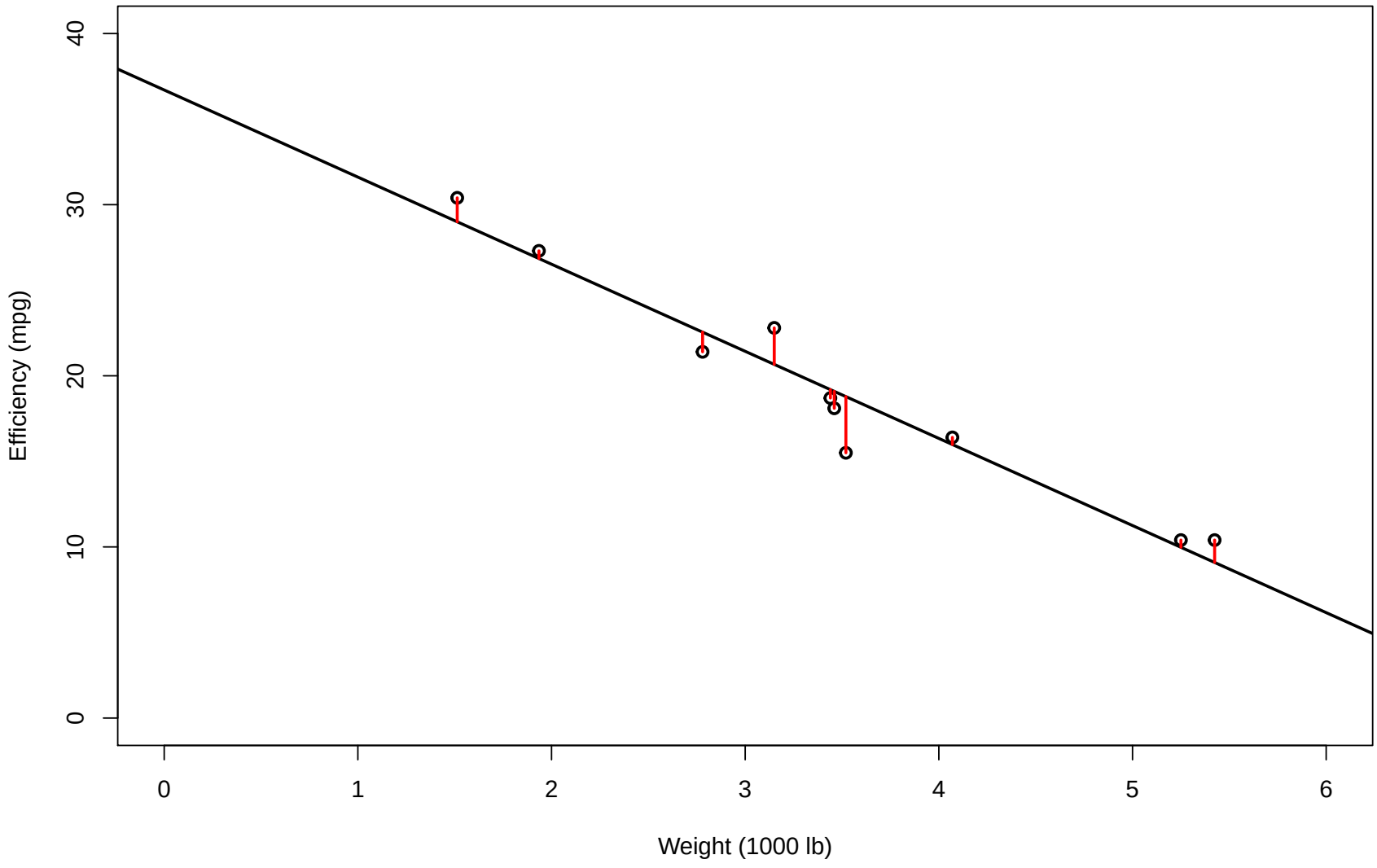
# Calculate and interpret the coefficient of determination

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

$$\text{coefficient of determination} = r^2 = (-0.971)^2 = 0.943$$

We say that 94.3% of the variance in $y$ (fuel efficiency) is explained by $x$ (weight).

# Determine the residuals (error between model and actual)



# Determine the residuals and standard error of regression

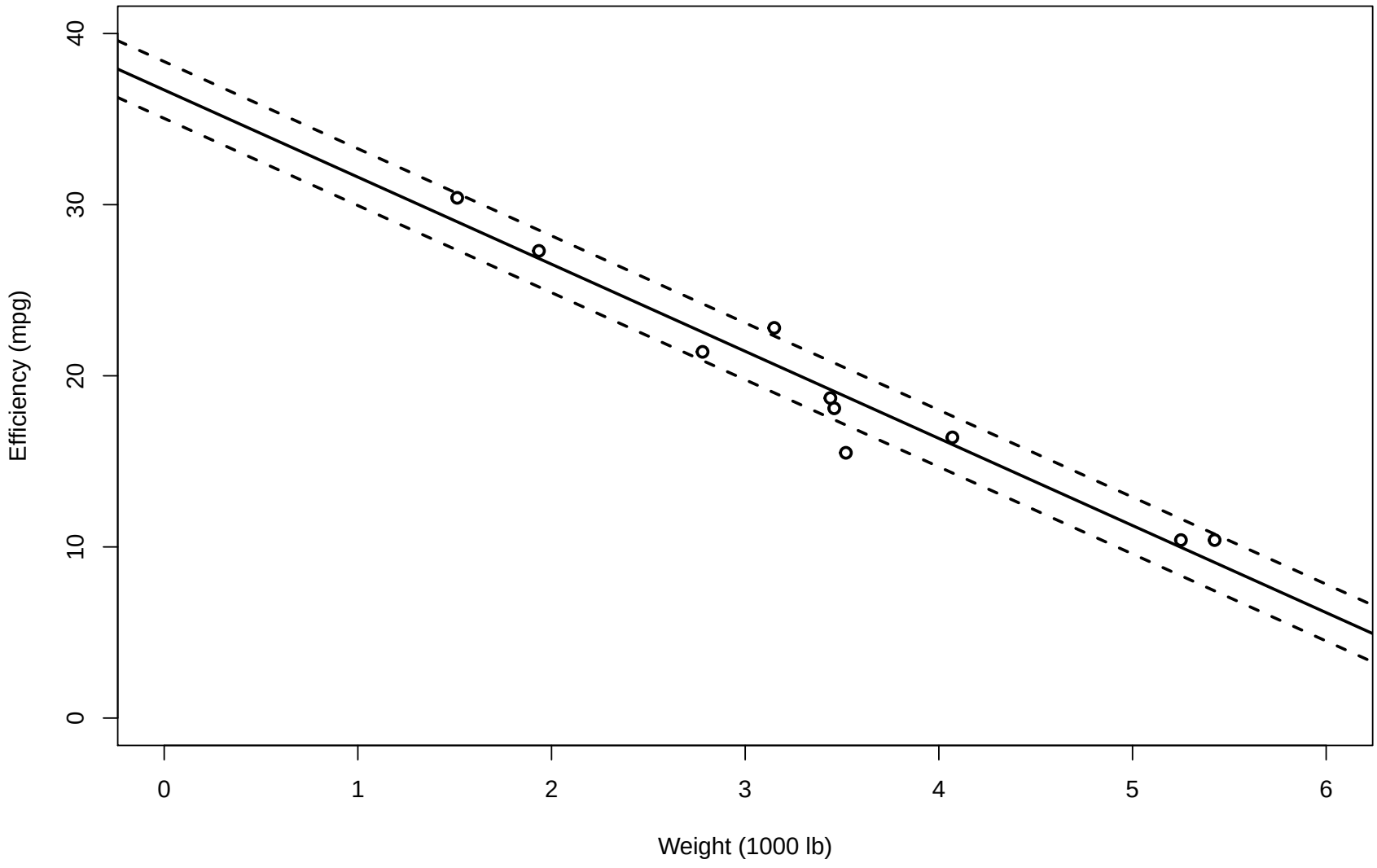
Let $y'$ be the predicted value from $y = a + bx$ linear model.

| $x$ | $y$ | $y'$ | $y - y'$ | $(y - y')^2$ |
|---|---|---|---|---|
| 1.513 | 30.4 | 29 | 1.4 | 1.96 |
| 5.424 | 10.4 | 9.09 | 1.31 | 1.72 |
| 1.935 | 27.3 | 26.9 | 0.4 | 0.16 |
| 3.52 | 15.5 | 18.8 | -3.3 | 10.9 |
| 3.44 | 18.7 | 19.2 | -0.5 | 0.25 |
| 4.07 | 16.4 | 16 | 0.4 | 0.16 |
| 5.25 | 10.4 | 9.98 | 0.42 | 0.176 |
| 3.15 | 22.8 | 20.7 | 2.1 | 4.41 |
| 2.78 | 21.4 | 22.5 | -1.1 | 1.21 |
| 3.46 | 18.1 | 19.1 | -1 | 1 |

$$s_{y-y'} = \sqrt{\frac{\sum (y - y')^2}{n - 2}} = \sqrt{\frac{21.946}{10 - 2}} = 1.66$$

We think predictions tend to be off by 1.66 mpg.

# Draw standard error



# Make a point estimate using the linear model

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

$$y = 36.7 + (-5.09)(0.2) = 35.682$$

. . . 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 200 mpg on an energy basis).