3.3 Finding a Regression Line

Dr. Lauren Perry

Goals

1. Calculate and interpret a regression line.
2. Interpret a coefficient of determination.
3. Understand the relationship between a correlation coefficient and a coefficient of determination.
4. Use a regression line to make predictions.
   - Identify potential problems with predictions.

Residuals

Residuals are the leftover *stuff* (variation) in the data after accounting for model fit:

$$\text{data} = \text{prediction} + \text{residual}$$

Residuals

- Each observation has its own residual.
- The residual for an observation (x, y) is the difference between observed (y) and predicted $(\hat{y})$:

$$e = y - \hat{y}$$

- We denote the residuals by e and find $\hat{y}$ by plugging x into the regression equation.

Note: If an observation lands above the regression line, $e > 0$. If below, $e < 0$.

Residuals

Goal: get each residual as close to 0 as possible.

To shrink the residuals toward 0, we minimize:

$$\begin{aligned}\sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2\end{aligned}$$

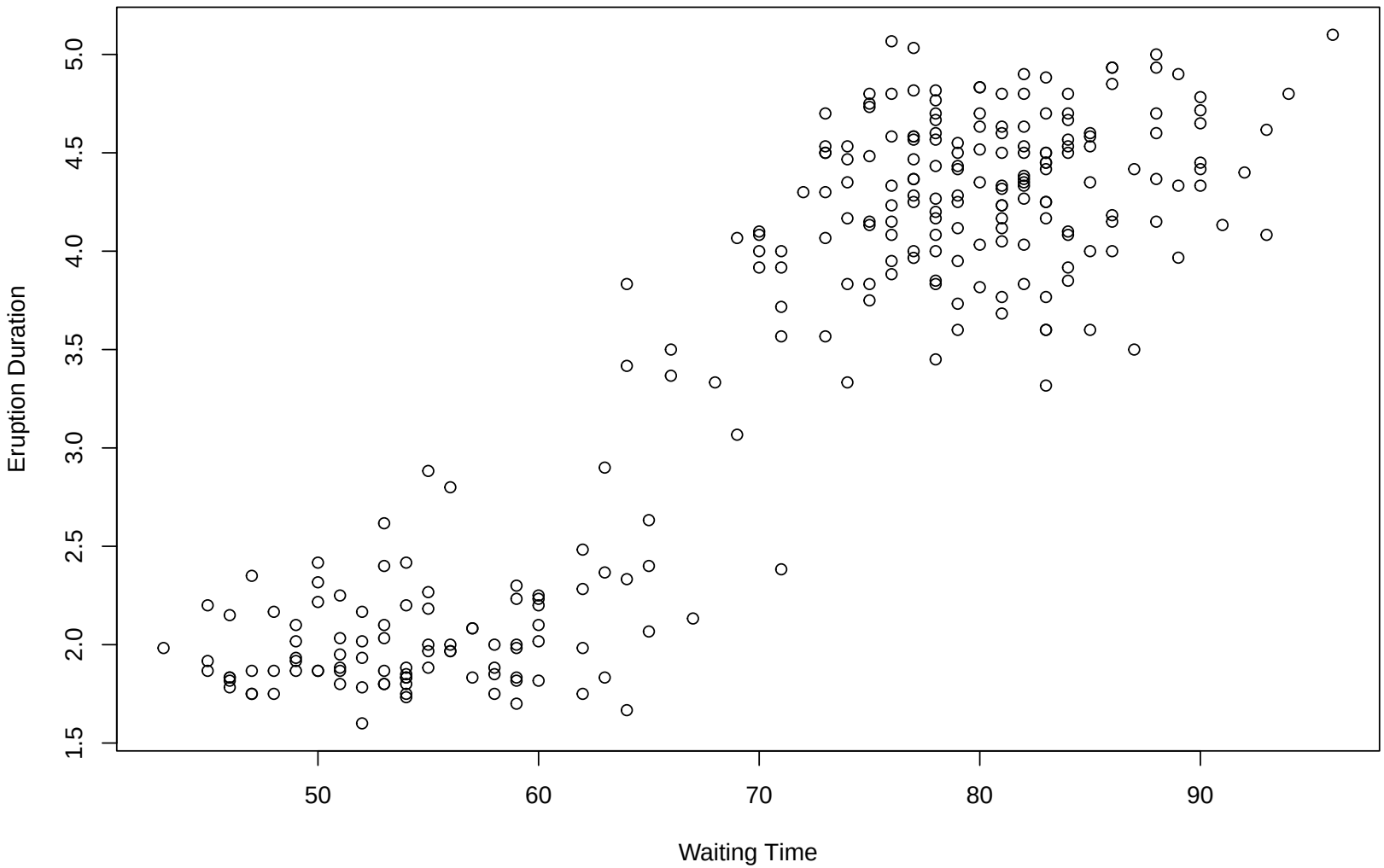
The values b_0 and b_1 that minimize this will make up our regression line.

Finding b_0 and b_1

- The slope is
$$b_1 = \frac{s_y}{s_x} \times R$$
- The intercept is
$$b_0 = \bar{y} - b_1 \bar{x}$$

Example: Old Faithful Geyser in Yellowstone

- `eruptions`, the length of each eruption
- `waiting`, the time between eruptions



Example: Old Faithful Geyser in Yellowstone

The sample statistics for these data are

	`waiting`	`eruptions`
mean	$\bar{x} = 70.90$	$\bar{y} = 3.49$
sd	$s_x = 13.60$	$s_y = 1.14$
		$R = 0.90$

Find the regression line and interpret the parameters.

Example: Old Faithful Geyser in Yellowstone

	`waiting`	`eruptions`
mean	$\bar{x} = 70.90$	$\bar{y} = 3.49$
sd	$s_x = 13.60$	$s_y = 1.14$
		$R = 0.90$

The equation for slope is

$$b_1 = \frac{s_y}{s_x} \times R$$

so

$$b_1 = \frac{1.14}{13.60} \times 0.90 = 0.075$$

Example: Old Faithful Geyser in Yellowstone

	`waiting`	`eruptions`
mean	$\bar{x} = 70.90$	$\bar{y} = 3.49$
sd	$s_x = 13.60$	$s_y = 1.14$
		$R = 0.90$

The equation for the intercept is

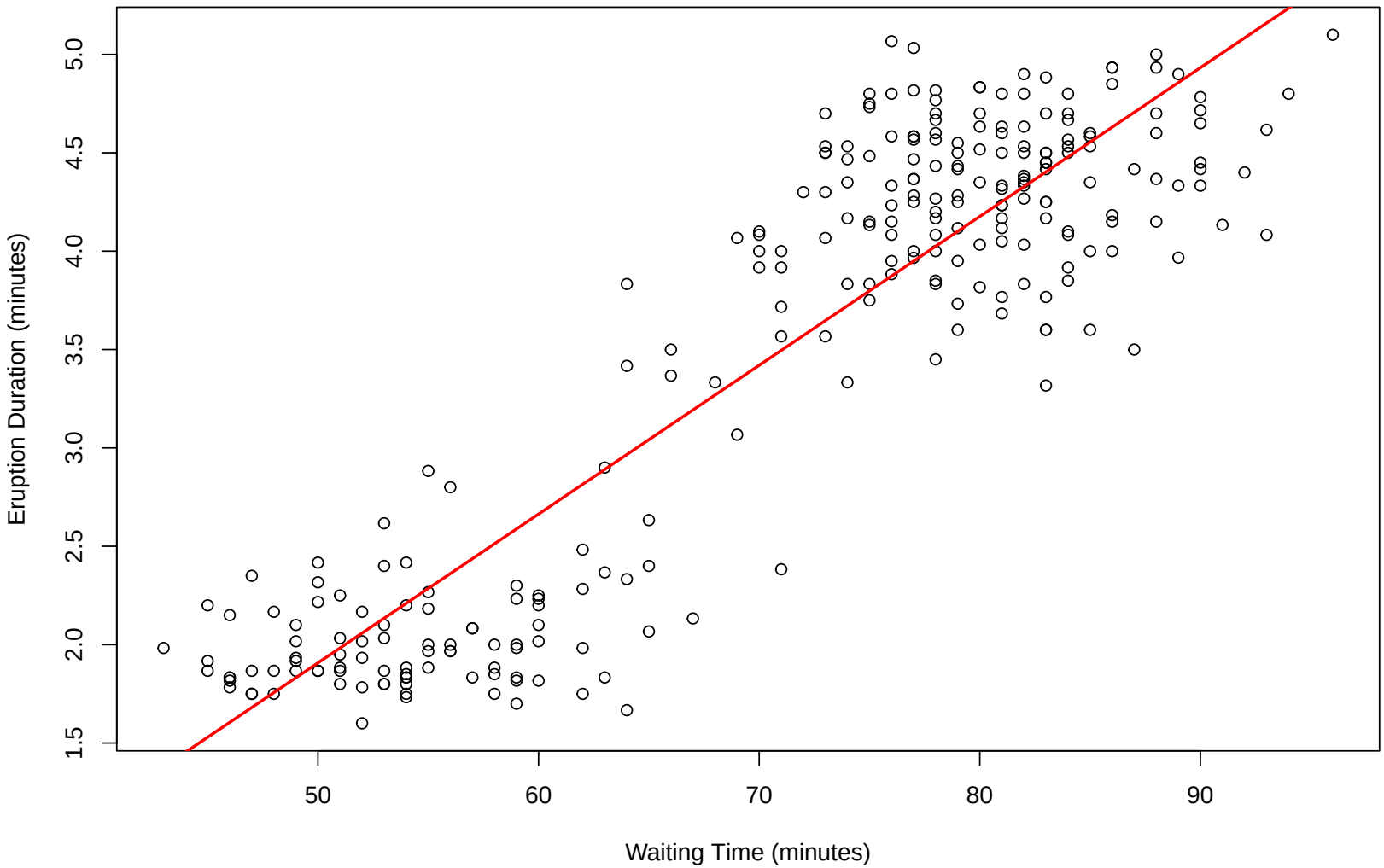
$$b_0 = \bar{y} - b_1 \bar{x}$$

and we found $b_1 = 0.093$ so

$$b_0 = 3.49 - 0.075 \times 70.90 = -1.83$$

Putting these together, the regression line is

$$\hat{y} = -1.83 + 0.075x$$



The Coefficient of Determination

The **coefficient of determination**, R^2, is the square of the correlation coefficient.

- This value tells us how much of the variability around the regression line is accounted for by the regression.
- An easy way to interpret this value is to assign it a letter grade.
  - For example, if $R^2 = 0.84$, the predictive capabilities of the regression line get a B.

Example: Old Faithful Geyser in Yellowstone

	`waiting`	`eruptions`
mean	$\bar{x} = 70.90$	$\bar{y} = 3.49$
sd	$s_x = 13.60$	$s_y = 1.14$
		$R = 0.90$

Since $R = 0.90$, we can find $R^2 = 0.90^2 = 0.81$

That is, 81% of the variability around the regression line is accounted for by the regression and this line gets a B

Prediction: Some Notes

It's important to stop and think about our predictions.

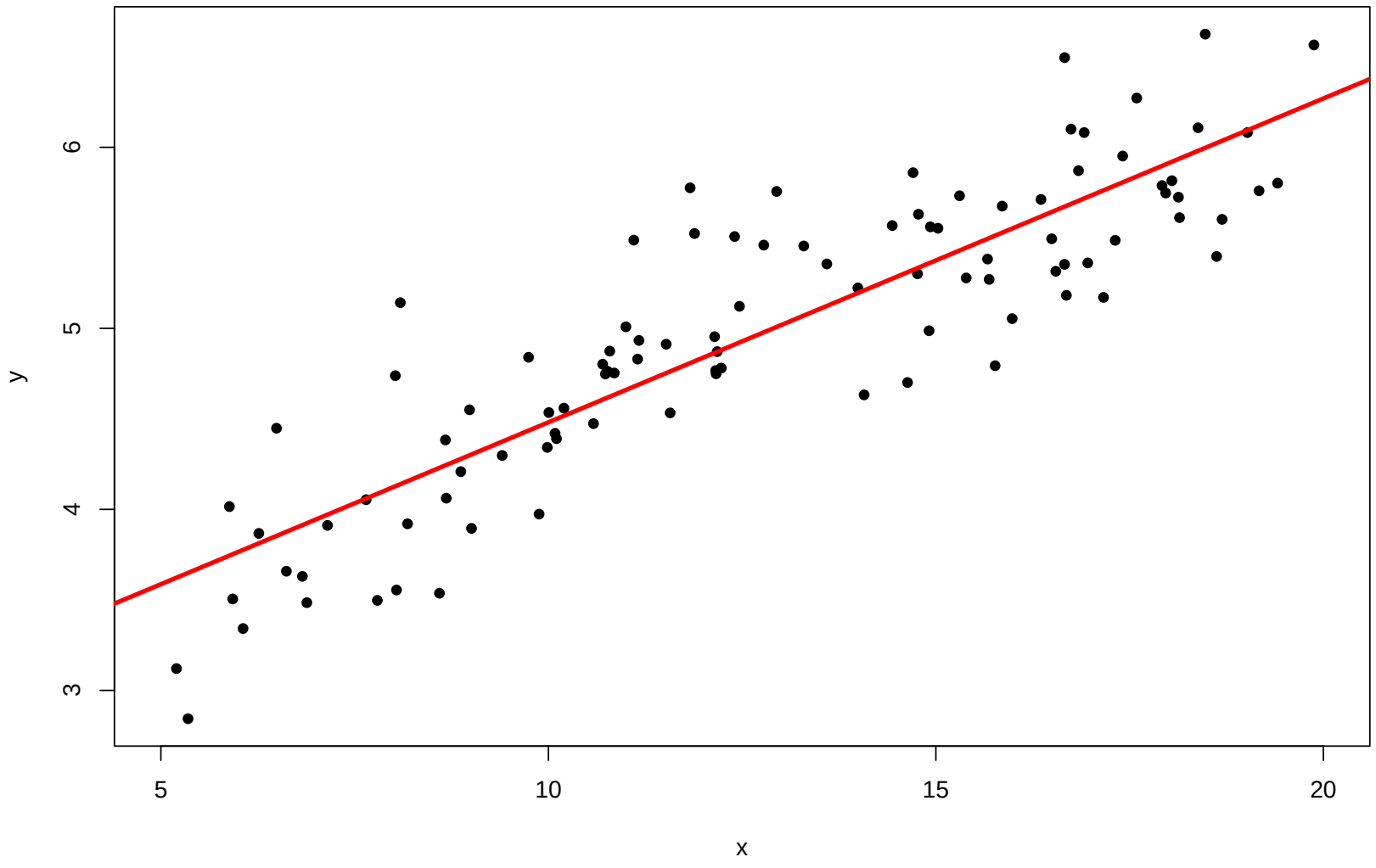
- Sometimes, the numbers just don't make sense.
- Other times it's harder to tell something's wrong!

Extrapolation

Extrapolation is applying a model estimate for values outside of the data's range for x.

- Our linear model is only an approximation.
  - We don't know anything about the relationship outside of the scope of our data.

Example



Example

The best fit line is

$$\hat{y} = 2.69 + 0.179x$$

- The correlation is $R = 0.877$.
- So the coefficient of determination is $R^2 = 0.767$.
  - (think: a C grade)

Prediction: Some Notes

Suppose we wanted to predict the value of y when $x = 0.1$:

$$\hat{y} = 2.66 + 0.181 \times 0.1 = 2.67$$

Prediction: Some Notes

This seems reasonable. . . but the true (population) best-fit model looks like this: