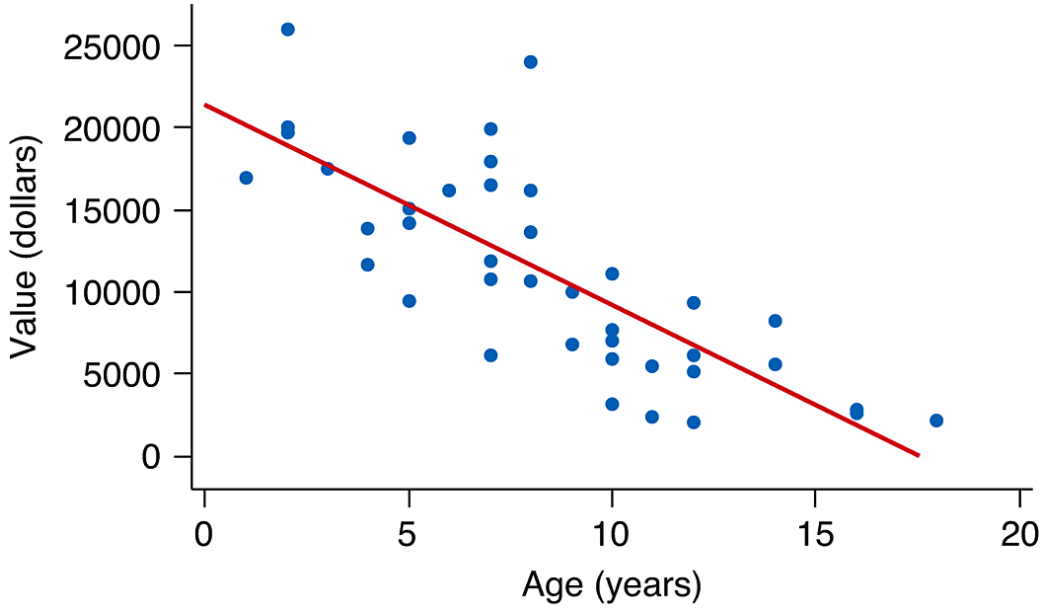
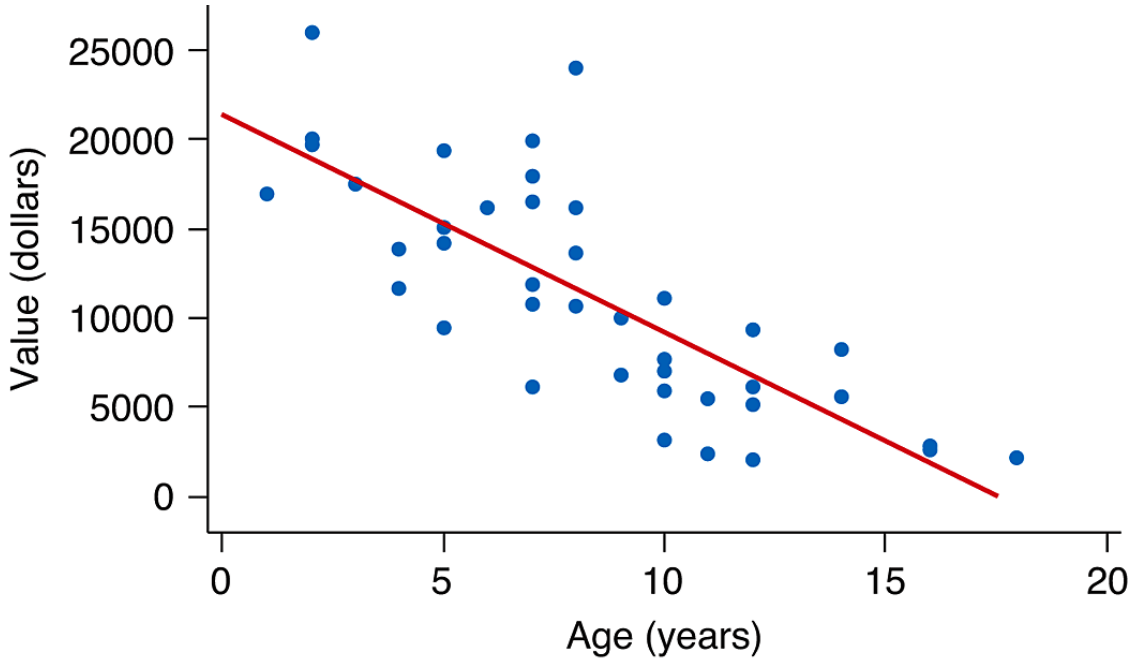
Section 4.3: Modeling Linear Trends



- The regression line is the line that best fits the data. It has the form $\boldsymbol{y = a + bx}$, where a is called the y-intercept and b is called the *slope*.
- Regression line is a tool to **summarize a linear relationship** and to **make predictions** about future observations.

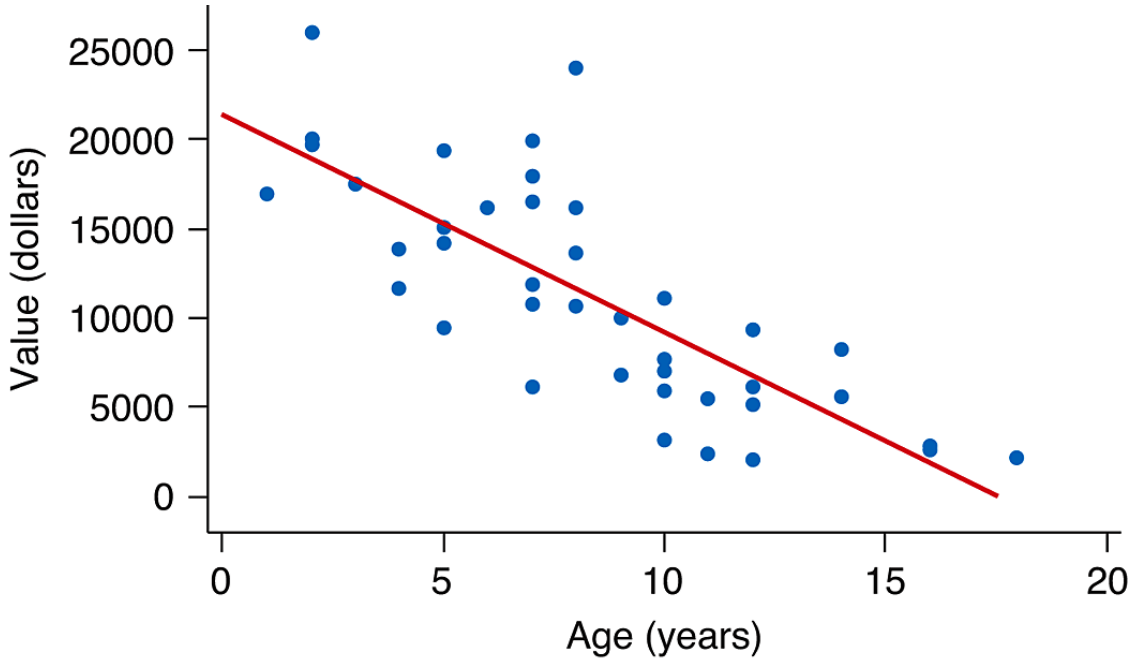
Example of using regression line



- The scatterplot shows a negative linear trend. As age of car increases, value tends to decrease.
- The regression equation is:

$$\text{predicted value} = 21375 - 1215 \times \text{age}$$

Example of using regression line



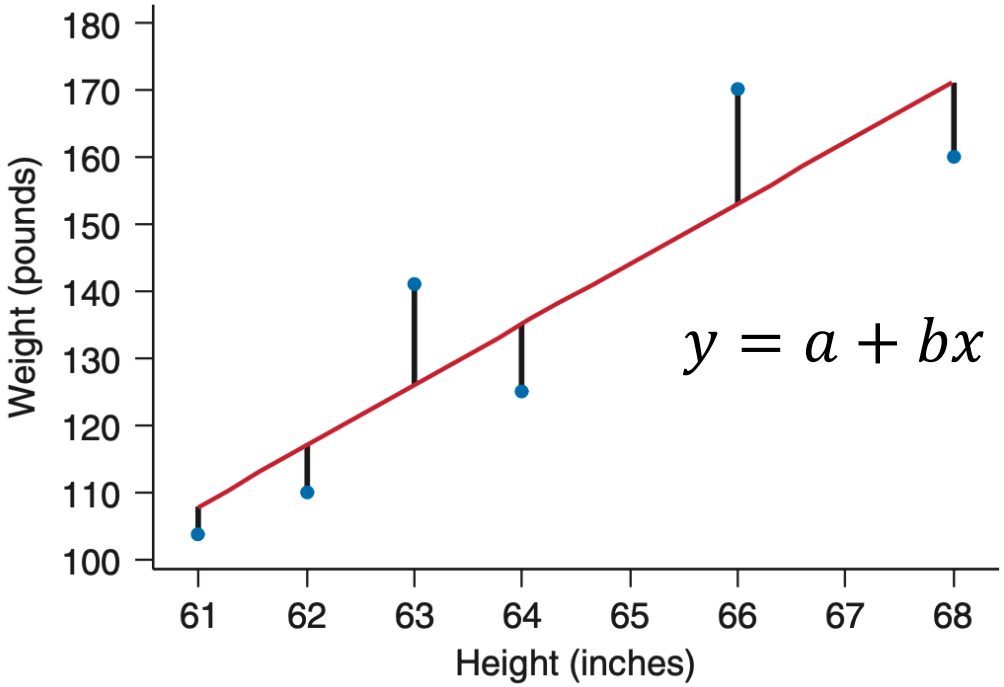
- The regression equation is:

$$\text{predicted value} = 21375 - 1215 \times \text{age}$$

- Predict the value of a 13-year-old car:

$$\text{predicted value} = 21375 - 1215 \times 13 = 5580$$

Finding the Regression Equation



a, b are chosen such that $\sum(y - a - bx)^2$ is minimum.

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}, \qquad a = \frac{\sum y}{n} - b\frac{\sum x}{n}$$

Example

x	1	3	4
y	2	2	4

$$\sum x = 1 + 3 + 4 = 8, \qquad \sum x^2 = 1^2 + 3^2 + 4^2 = 26$$

$$\sum y = 2 + 2 + 4 = 8, \qquad \sum y^2 = 2^2 + 2^2 + 4^2 = 24$$

$$\sum xy = 1(2) + 3(2) + 4(4) = 24$$

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2} = \frac{3 \times 24 - 8 \times 8}{3 \times 26 - 8^2} = \frac{4}{7}$$

$$a = \frac{\sum y}{n} - b\frac{\sum x}{n} = \frac{8}{3} - \frac{4}{7} \times \frac{8}{3} = \frac{8}{7} \qquad \boldsymbol{y = \frac{8}{7} + \frac{4}{7}x}$$

Using Stat Crunch

Enter the data into StatCrunch.

Stat > Regression > Simple Linear

Select the *x*-variable, select the *y*-variable, select **Compute.**

Interpreting the Regression Line

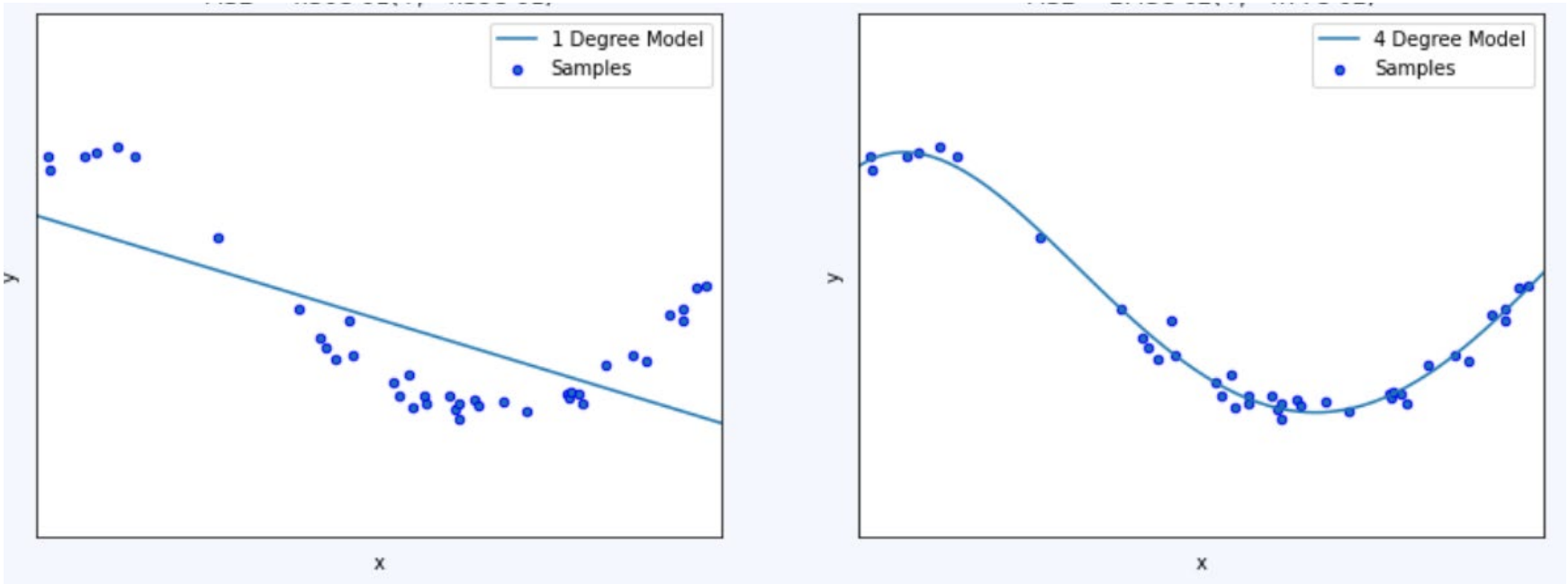
- Switch *x* and *y* will change the regression equation.
- We use the *x*-variable to make predictions about the *y*-variable, so the *x*-variable is called the *explanatory* or *predictor* or *independent* variable.
- The *y*-variable is the response or predicted variable. It is called the *dependent* variable.
- The slope b has the same sign as the correrlation coefficient r.
  - $b > 0$: positive correlation
  - $b < 0$: negative correlation
  - $b = 0$: no correlation

Section 4.4
Evaluating the Linear Model

- When to use linear models to describe relationships?
- What are some pitfalls to avoid?

Use linear models to when appropriate

- Don’t use linear models to describe non-linear associations. Always look at a scatterplot first.



Coefficient of Determination: r^2

- Coefficient of determination is the square of *r*, the correlation coefficient.
- Because $-1 \leq r \leq 1$, we have $0 \leq r^2 \leq 1$. In practice, r^2 is usually converted to a percentage.
- r^2 measures how much variation in the response variable is explained by the explanatory variable.
- The larger r^2, the smaller the amount of variation or scatter about the regression line, and the more accurate the predictions tend to be.

Example: r^2

For the data on car age and predicted value, $r = -0.778$.

$$r^2 = (-0.778)^2 = .605, \text{ so } r^2 = 60.5\%.$$

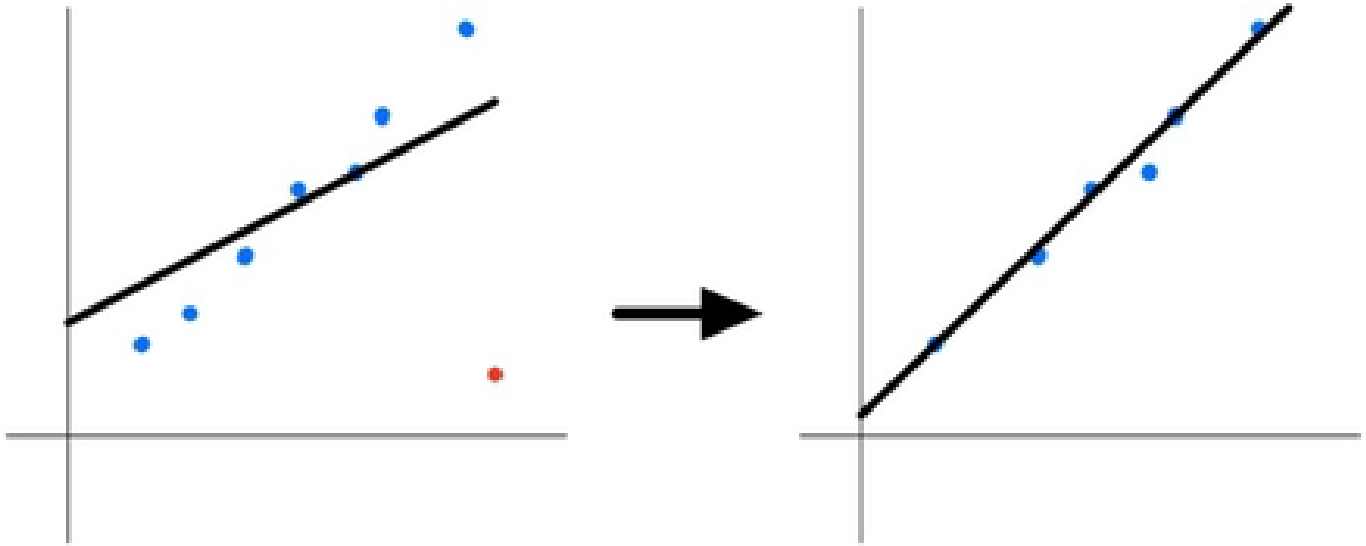
Car age explains about 60.5% of the variation in car value.

Cause and effect

A correlation between two variables is not sufficient evidence to conclude that a cause-and-effect relationship exists between the variables, no matter how strong the correlation or how well the regression line fits the data.

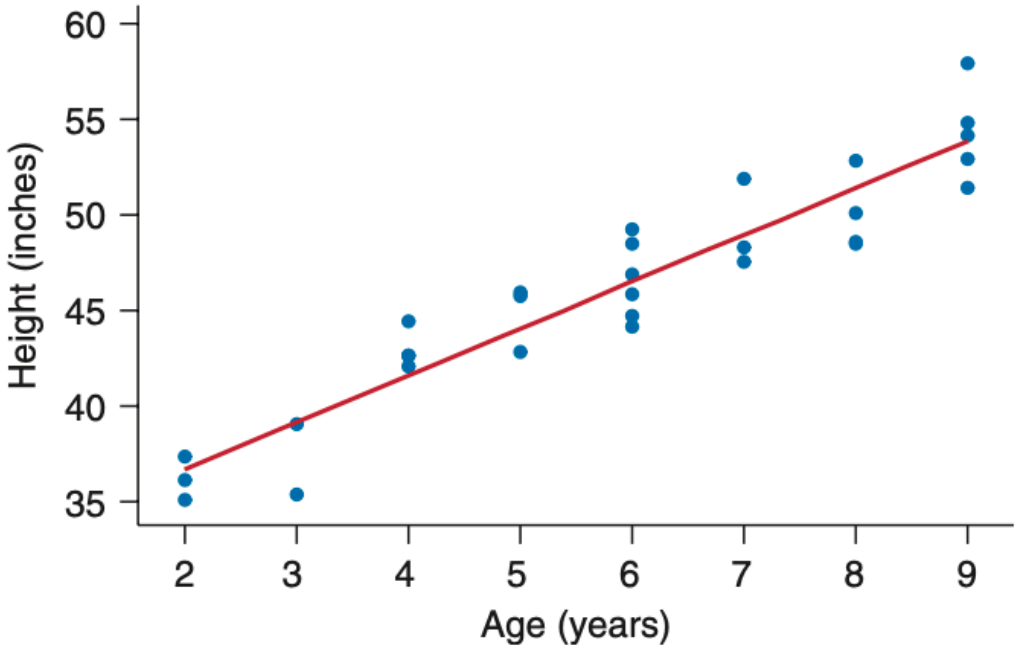
Remove influential points if necessary

Influential points are data points that, if removed, would significantly change the slope or intercept of the regression line.



Avoid extrapolation

Don’t make predictions beyond the range of the data, because we are not sure that the linear trend will continue beyond the range of the data.



You can’t use this linear model to predict the height of a 20-year-old person.