



CONTENTS

1	Manufacturing	3
1.1	Woods and Metals Processes	3
1.1.1	Milling	3
1.1.2	Lathes	5
1.2	Metal-specific Processes	6
1.2.1	Sheet Metal Processes	6
1.2.2	Casting	7
1.3	Wood-specific Processes	8
1.3.1	Bending	8
1.4	Plastic-specific Processes	9
1.4.1	Injection Molding	9
1.5	3D Printing	10
1.6	Laser Cutting and Water Jet	11
2	Introduction to Waves	15
2.1	Basic Properties of Waves	15
2.2	Two Main Types of Waves	17
2.3	Compression and Rarefaction	18
2.4	Creating a Wave in Python	18
3	Sound	21
3.1	Pitch and frequency	22
3.2	Chords and harmonics	24
3.3	Making waves in Python	25

3.3.1	Making a sound file	27
4	Scatter Plots and Correlation	31
4.1	Review of Scatter Plots	31
4.2	Correlation Coefficient	31
5	The Regression Line	39
5.1	Linear Regression	39
6	Data Analysis with jamovi	51
6.1	Getting Started with jamovi	51
6.2	Setting Up Your Data	52
6.3	Describing Data	53
6.4	Comparing Two Groups	54
6.5	Looking at Relationships	55
6.6	Saving Your Work	55
7	Transforming Nonlinear Data with Logarithms	57
7.1	Exponential Models	57
7.2	Power Models	61
7.3	Choosing the Right Transformation	64
A	Answers to Exercises	67
	Index	75

Manufacturing

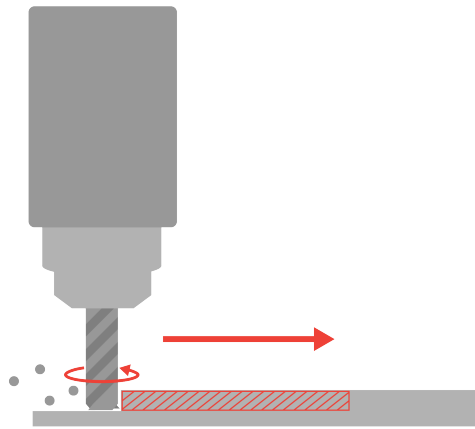
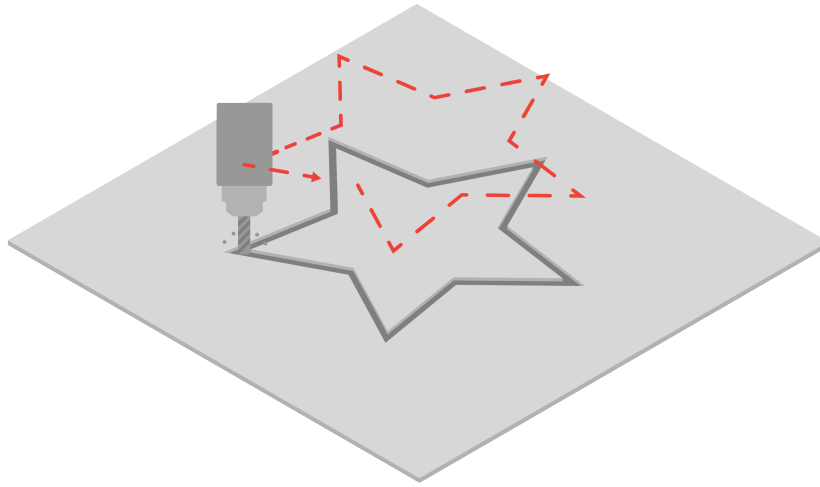
Look around the room. Your chair, your phone case, the hinges on the door, the pen in your hand: every one of them was made by a machine following a process that engineers designed. Why is a phone case plastic and molded, while a bike frame is metal and welded? Each manufacturing process has its strengths, and choosing the right one means balancing cost, quantity, precision, and time. Here is a tour of the most important ones.

1.1 Woods and Metals Processes

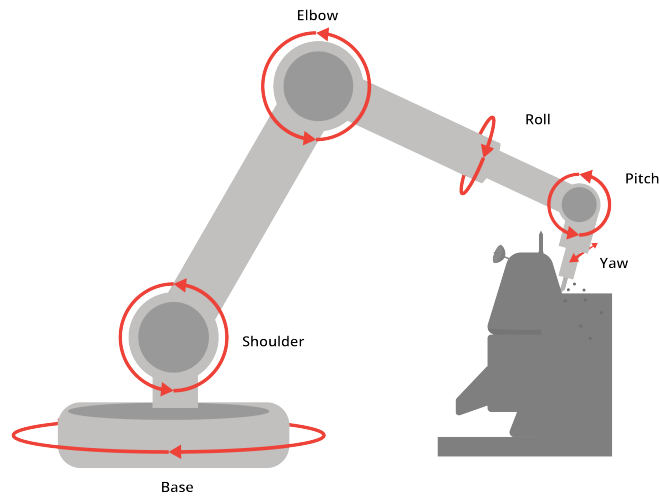
Bridges, cars, buildings, and furniture all depend on shaping wood and metal accurately. These first two machines work on both.

1.1.1 Milling

A mill carves shapes out of a solid block. A spinning cutting tool follows a path and removes material, so milling is a *subtractive* process: you start with more than you need and cut away the rest, like a sculptor. The tool spins very fast, which lets it cut hard materials like wood and metal.



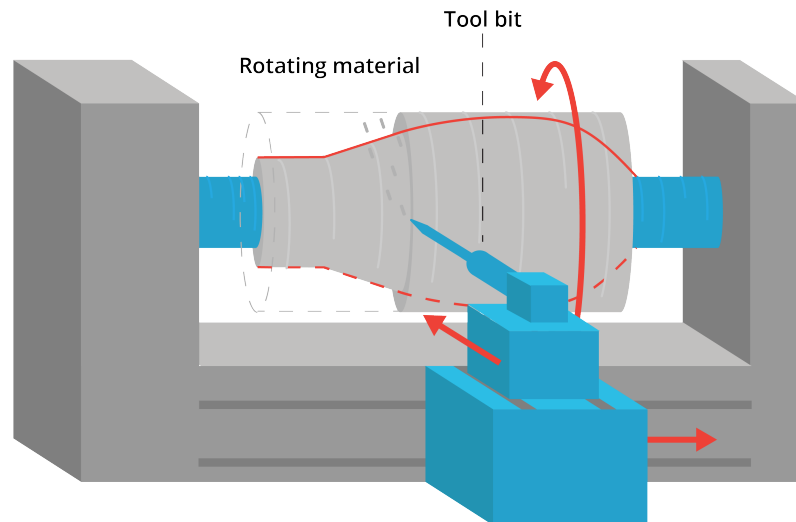
A 3-axis mill moves the tool along the x , y , and z axes, which is enough for many parts. A 5-axis (or even 6-axis) mill can also tilt and rotate, so it can reach around curves and undercuts, like the blades of a jet engine turbine.



In manufacturing settings where speed and repeatability is paramount, mills are often computer controlled. This functionality is referred to as *Computer Numerical Control*, or CNC. CNC mills are able to repeatedly follow a tool path, resulting in consistent and accurate parts.

1.1.2 Lathes

A lathe makes round things: bolts, axles, table legs, baseball bats. It is also subtractive, but it flips the mill's idea around. The *material* spins, and a stationary cutting tool shaves it down, layer by layer. (Shaping on a lathe is called *turning*.)



In the manufacturing industry, lathes are also often computer controlled. Alongside CNC mills, these CNC lathes are responsible for a majority of the objects that we interact with everyday.

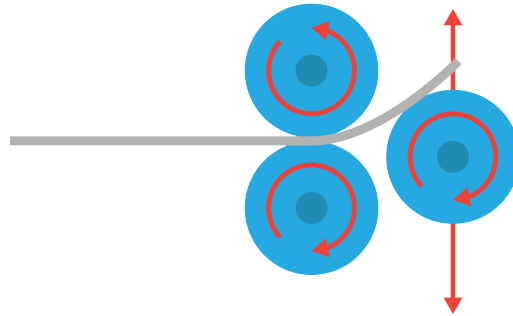
1.2 Metal-specific Processes

While mills and lathes are already able to cover the vast majority of manufacturing needs for woods and metals, there are certain processes that are specifically enabled by the unique properties of metals. Most of them take advantage of malleability: metals can be pressed, rolled, and bent into new shapes without cracking, especially when warm.

1.2.1 Sheet Metal Processes

While milling allows us to process blocks of metal to great effect, sometimes the application we need does not require material of such thickness.

This is where sheet metal comes in, as well as the methods we use to process it. One of the most common techniques in manufacturing is rolling, where a raw sheet is gradually rolled into the desired shape. This method is used to create many objects you may recognize, such as metal roofing, the skin of an airplane, and soda cans.



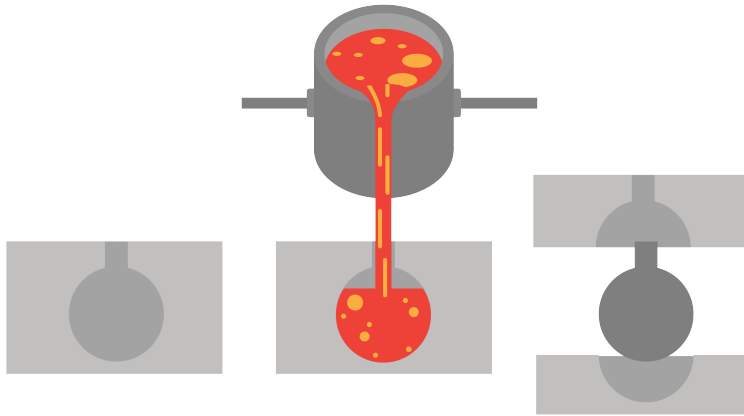
Another method is stamping, where a raw sheet is stamped into the desired shape. This allows us to create surfaces with complex geometries in an instant, and in large quantities. This method is often used for applications like the exterior panels of a car, where parts with compound curves are needed.



Finally, sheet metal parts are often stiffened by bending their edges over into a lip, called a *flange* or a *hem*. Look at the edge of a metal shelf or a car door: the folded edge makes the part much stiffer, and it hides the sharp cut edge.

1.2.2 Casting

The last kind of metal-specific process we will cover is casting. Casting involves pouring molten metal into a mold, then letting the metal cool and set inside the mold. Casting can make complex shapes in large numbers: engine blocks, faucets, and die-cast toy cars are all cast.



Casting also results in minimal material wastage, as it is not a subtractive manufacturing method where excess material is machined away, but rather only the specific amount of material required is poured in each time.

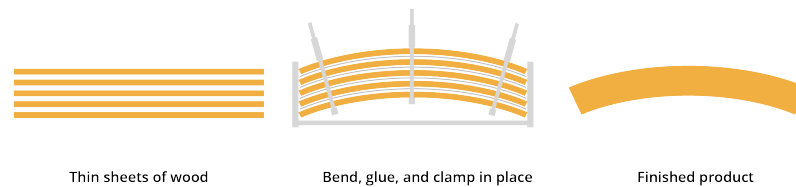
1.3 Wood-specific Processes

Wood is made of long fibers, so it is strong along the grain and flexible when thin. Woodworkers use that to make curves.

1.3.1 Bending

Turning raw wood into flat, workable pieces involves a variety of tools that you probably know of already, such as saws and drills. However, there are specific processes that help us create curved shapes with wood, in addition to the mill and the lathe mentioned earlier.

This is where bent lamination comes in. Bent lamination involves layering multiple thin veneers or strips of wood with adhesive, and clamping it to create the desired form while letting the glue dry. This method is often used for furniture production, enabling continuous curves in wood with tight radii.



Steam bending works differently. Hot steam softens the natural glue (lignin) that holds wood fibers together, so a solid piece of wood can be bent around a form. As it cools and dries, it holds the new shape. Unlike bent lamination, steam bending needs no adhesive. Wooden boat ribs and classic bentwood chairs are made this way.

1.4 Plastic-specific Processes

Although plastics only came into prominence in the mid 20th century, they have changed manufacturing and, by extension, the world as we know it. Easily manufacturable, durable, and cost-effective, they have come to permeate almost everything we use on a daily basis.

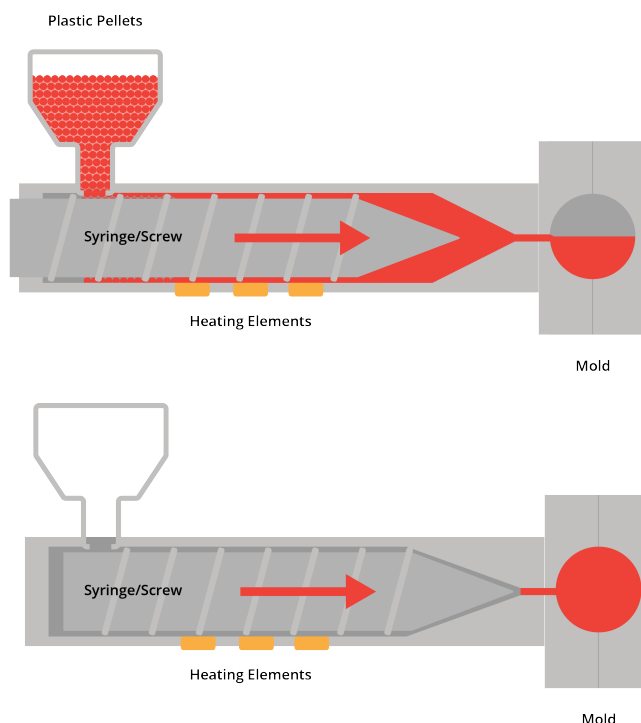
It must also be noted that these exact qualities have also resulted in the proliferation of plastics in our environment, and as such, usage of plastics should be well considered and limited. Alternative biodegradable materials are currently being trialed by scientists and would look to replicate many of the same qualities we expect from plastics, including its manufacturability.

1.4.1 Injection Molding

Injection molding is responsible for the vast majority of plastic products that you interact with on a daily basis. It is extremely quick, highly accurate, and has minimal material wastage, making it a popular and cost effective method of manufacturing plastic goods.

Similar to casting, injection molding involves injecting molten plastic into a mold, then allowing the part to cool and set inside the cavity.

You can often tell when a part was produced through injection molding, with telltale signs such as the parting line. This is where the parts of the mold meet, forming a visible line on the surface of a part.



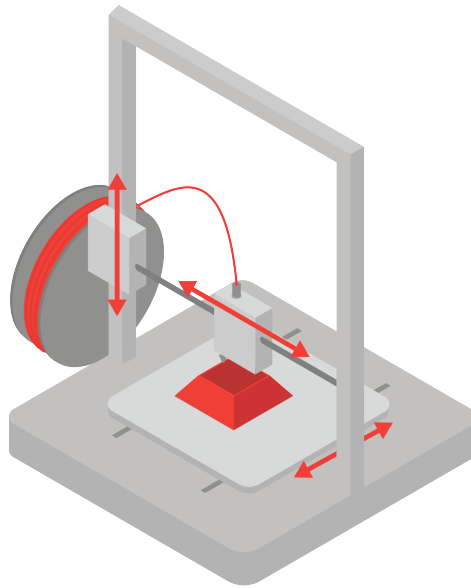
1.5 3D Printing

Injection molding has traditionally been the go-to technique for manufacturing plastic goods. However, new technologies result in the advent of new manufacturing methods. 3D printing has become popular in the last few decades as a way to make a part quickly without first building an expensive mold.

3D printing is an *additive* process: it builds a part one thin layer at a time. The most common printers melt plastic filament and lay it down like a very precise hot glue gun. Others harden liquid resin with light or fuse metal powder with a laser. 3D printers can make shapes no mold could, like a part with hollow channels inside. They are slower and usually less precise than injection molding, but they are improving rapidly.

Worked example. A part 30 mm tall is printed with 0.2 mm layers. How many layers does it need?

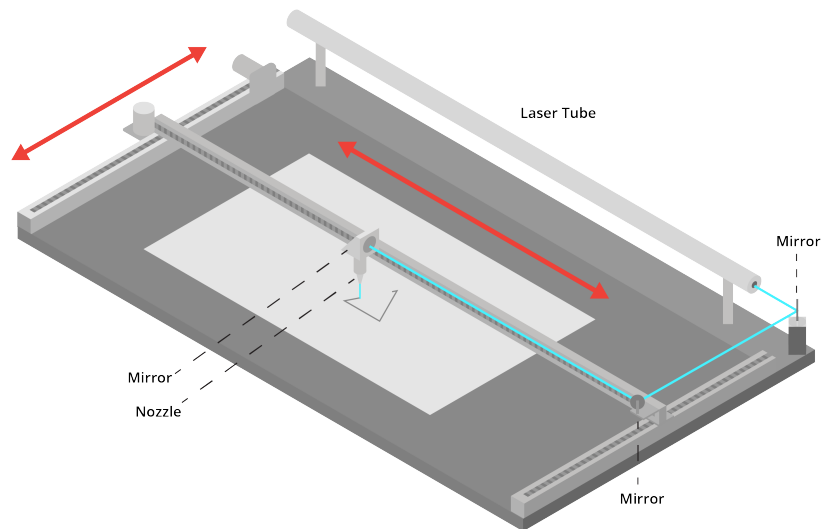
Solution. $\frac{30}{0.2} = 150$ layers.



1.6 Laser Cutting and Water Jet

A laser cutter focuses a powerful beam of light to a tiny spot that burns or melts its way through sheet material as the head moves. Small laser cutters in school makerspaces cut and engrave wood, acrylic, cardboard, and fabric; big industrial lasers slice sheet steel. Laser cutting is subtractive, and it works best on thin sheets.

A water jet cutter blasts water through a tiny nozzle at several times the speed of sound. Mixed with grit, the stream cuts metal, stone, glass, and rubber. Unlike a laser, it does not heat the material, so it will not warp or scorch the edge. Very thick material is still hard to cut.



Exercise 1 More Layers

A part 45 mm tall is printed with 0.1 mm layers. How many layers are needed? If each layer takes 30 seconds, how long does the print take?

Working Space

Answer on Page 67

Exercise 2 Mold or printer?

You need plastic phone stands. An injection mold costs \$20,000 to make, and then each stand costs \$0.50. A 3D-printing service charges \$8.00 per stand and needs no mold. For how many stands is injection molding the cheaper choice? Which would you use for 100

Working Space

stands? For 100,000?

Answer on Page 67

Introduction to Waves

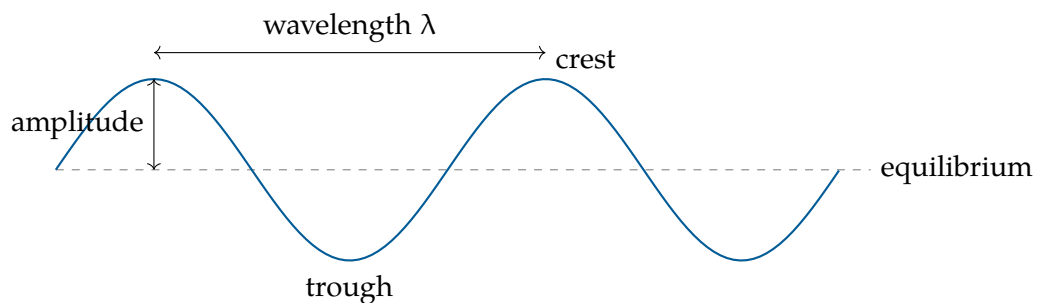
Have you ever done “the wave” at a stadium? Each fan stands up and sits down, never leaving their seat, yet a ripple races all the way around the stadium. That is the essence of every wave in nature. Ripples on a pond, the sound from a speaker, the light from the Sun, the radio signal reaching your phone, the shaking of an earthquake: all of them are waves, and all of them share the same basic properties. Learn those properties once, and you can understand them all.

A **wave** is a disturbance that carries energy from one place to another without carrying matter along with it.

When a ripple crosses a pond, the water does not travel across the pond. A floating leaf just bobs up and down as the ripple passes. The water *oscillates* around its resting position while the wave carries energy outward, just like the fans in the stadium.

2.1 Basic Properties of Waves

The simplest waves have the shape of the sine curves you studied in the trigonometry chapters. Five terms describe them.



Equilibrium

The *equilibrium* position is where the medium sits when there is no wave: the flat surface of the pond before the pebble lands, or a rope lying still.

Amplitude

The *amplitude* is the largest distance the medium moves away from equilibrium. The highest points of the wave are *crests* (or peaks), and the lowest points are *troughs* (or valleys). In a simple wave, the crests are as far above equilibrium as the troughs are below it.

For a water wave, the amplitude is the height of a crest above the resting water level. A bigger amplitude means more energy: louder sound, brighter light, a more violent earthquake.

Wavelength

The *wavelength*, written λ (the Greek letter lambda), is the distance from one point on the wave to the matching point on the next wave, such as crest to crest. It is measured in meters.

Frequency

The *frequency* f is the number of complete waves that pass a point each second. It is measured in hertz (Hz): 1 Hz = 1 wave per second. For sound, higher frequency means higher pitch. For light, frequency determines color.

Period

The *period* T is the time for one complete oscillation. If 4 waves pass each second, each one takes $\frac{1}{4}$ of a second. In general,

$$T = \frac{1}{f}$$

Wave Speed

In one period, the wave moves forward exactly one wavelength. Since speed is distance divided by time,

$$v = \frac{\lambda}{T} = f\lambda$$

The speed depends on the medium. Sound travels about 343 m/s in air and about 1,480 m/s in water. Light travels 300,000 km/s through empty space.

Worked example. The note A above middle C has a frequency of 440 Hz. Its wavelength in air is 0.78 m. How fast does it travel, and what is its period?

Solution. $v = f\lambda = (440)(0.78) \approx 343$ m/s. The period is $T = \frac{1}{440} \approx 0.00227$ s, about 2.3 milliseconds.

Exercise 3 Label the wave

A wave on a rope is drawn on a grid. The rope's resting position is $y = 0$. Its crests reach $y = 15$ cm and its troughs reach $y = -15$ cm. Crests are 80 cm apart, and 5 crests pass a fixed point every 2 seconds. Find the amplitude, wavelength, frequency, period, and speed.

Working Space

Answer on Page 67

Exercise 4 Wavelength of a Low Note

A sound wave has frequency 20 Hz (about the lowest pitch humans can hear) and travels at 343 m/s. What is its wavelength?

Working Space

Answer on Page 67

2.2 Two Main Types of Waves

Waves come in two kinds, depending on which way the medium moves as the wave passes.

Transverse Waves

In a *transverse* wave, the medium moves *perpendicular* to the direction the wave travels. Flick one end of a rope tied to a wall: the rope moves up and down, while the wave runs along it toward the wall. The stadium wave is transverse, too. Light and other electromagnetic waves are transverse.

Longitudinal Waves

In a *longitudinal* wave, the medium moves back and forth *parallel* to the direction the wave travels. Sound is the most important example. A speaker cone pushes on the air next to it, that air pushes on the air next to it, and so on, until the push reaches your ear. Each bit of air only jiggles back and forth.

Your digital resources include a video of both kinds of waves on a slinky. Push the end of a stretched slinky in and out, and you get a longitudinal wave; shake it side to side, and you get a transverse one.

2.3 Compression and Rarefaction

A longitudinal wave has no crests or troughs. Instead, it has regions where the medium is squeezed together, called *compressions*, and regions where it is spread apart, called *rarefactions*. In a sound wave, the air pressure is slightly higher in the compressions and slightly lower in the rarefactions.



The distance from one compression to the next is the wavelength. If you graph the air pressure along the wave, you get a sine curve, so all the properties from the last section still apply.

2.4 Creating a Wave in Python

A traveling wave is a sine curve that slides along as time passes. With amplitude A , wavelength λ , and frequency f , the height of the medium at position x and time t is

$$y = A \sin \left(2\pi \left(\frac{x}{\lambda} - ft \right) \right)$$

Let us draw the wave at a few moments to watch it move. Create `wave.py`:

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 A = 0.15          # amplitude (m)
5 wavelength = 0.8  # m
6 f = 2.5           # Hz
7
8 x = np.linspace(0, 2.4, 300)
9 for t in [0, 0.1, 0.2]:
10     y = A * np.sin(2 * np.pi * (x / wavelength - f * t))
11     plt.plot(x, y, label=f"t = {t} s")
12
13 plt.axhline(0, color="gray", linewidth=0.8)
14 plt.xlabel("position (m)")
15 plt.ylabel("displacement (m)")
16 plt.legend()
17 plt.show()
```

These are the numbers from the rope exercise. Each curve is the same sine wave, shifted to the right. In 0.1 s, a wave moving at 2 m/s should move 0.2 m, and that is exactly how far the crests shift between curves.

Exercise 5 Faster waves

Change `f` to 5 and run the program again, leaving the wavelength alone. How far do the crests move between curves now? What happened to the wave speed?

Working Space

Answer on Page 67

Exercise 6 Radio waves

An FM radio station broadcasts at 100 MHz (100 million hertz). Radio waves travel at the speed of light, 3×10^8 m/s. What is the wavelength? What is the pe-

Working Space

riod?

Answer on Page 68

CHAPTER 3

Sound

Bang! A firecracker goes off down the street, and a fraction of a second later you hear it. What actually traveled from the firecracker to your ear? Not air: the air near the firecracker never reaches you. What reached you was a wave.

Let us break that down. Inside the cardboard wrapper of the firecracker, there is potassium nitrate (KNO_3), sulfur (S), and carbon (C). These are all solids. When you trigger the chemical reactions with a little heat, these atoms rearrange themselves to be potassium carbonate (K_2CO_3), potassium sulfate (K_2SO_4), carbon dioxide (CO_2), and nitrogen (N_2). Note that the last two are gasses.

The molecules of a solid are much more tightly packed than the molecules of a gas. So after the chemical reaction, the molecules expand to fill a much bigger volume. The air molecules nearby get pushed away from the firecracker. They compress the molecules beyond them, and those compress the molecules beyond them.

This compression wave spreads out as a sphere whose radius grows at about 343 meters per second: the *speed of sound* in air.

The energy of the explosion is distributed around the surface of this sphere. As the radius increases, the energy is spread more and more thinly around. This is why the firecracker seems louder when you are closer to it. (If you set off a firecracker in a sewer pipe, the sound will travel much, much farther.)

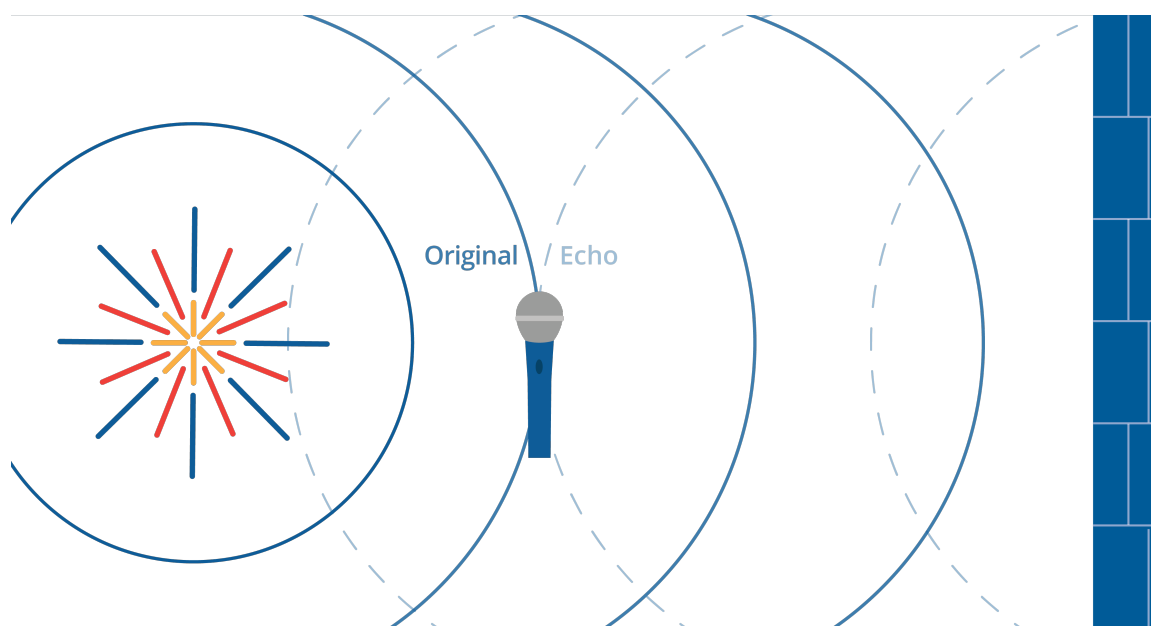


Figure 3.1: A firecracker exploding causes an initial sound wave and an echo.

This compression wave bounces off hard surfaces. Set off a firecracker 50 meters from a big wall, and you hear the explosion twice. The second bang, the *echo*, had to travel to the wall and back, 100 meters, so it arrives $\frac{100}{343} \approx 0.29$ seconds later.

The compression wave will be absorbed by soft surfaces. If you covered that wall with pillows, there would be almost no echo.

The study of how these compression waves move and bounce is called *acoustics*. Before you build a concert hall, you hire an acoustician to look at your plans and tell you how to make it sound better.

3.1 Pitch and frequency

A guitar string is a lot like the skate on a spring from the trigonometric functions chapter. The farther the string is displaced, the more force it feels pushing it back to equilibrium (remember the tension force?). Thus, it moves back and forth in a sine wave. (OK, it is not a pure sine wave, but we will get to that later.)

The string is attached through the bridge to the guitar's hollow wooden body, which the string pushes and pulls. That creates compression waves in the air around it.

If you are in the room with the guitar, those compression waves enter your ear and push and pull your eardrum, which is attached to bones that move a fluid that bends tiny hairs on the *hair cells* of your inner ear. This is how you hear.

We sometimes see plots of sound waveforms. The x-axis represents time. The y-axis represents the amount the air is compressed at the microphone that converted the air pressure into an electrical signal.

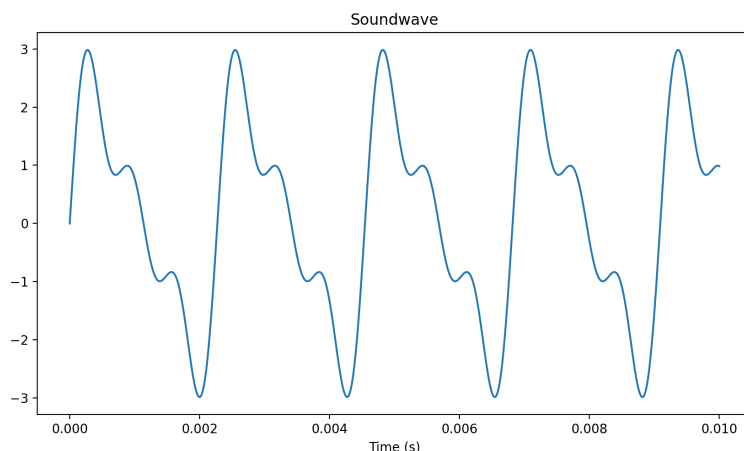


Figure 3.2: A soundwave graph over time.

If the guitar string is made tighter (by the tuning pegs) or shorter (by the guitarist's fingers on the strings), the string vibrates more times per second. We measure the number of waves per second and we call it the *frequency* of the tone. The unit for frequency is *Hertz*: cycles per second. The *period* is the reciprocal of the frequency: the time it takes for one cycle to complete. The unit for period is seconds per cycle.

Musicians have given the different frequencies names. If the guitarist plucks the lowest note on the guitar, it will vibrate at 82.4 Hertz. The guitarist will say "That pitch is low E." If the string is made half as long (by a finger on the 12th fret), the frequency will be twice as fast (164.8 Hertz), and the guitarist will say "That is E an octave up."

For any note, the note that has twice the frequency is one octave up. The note that has half the frequency is one octave down.

Worked example. The note A above middle C is 440 Hz. What are the A's one octave up and one octave down?

Solution. One octave up: $2 \times 440 = 880$ Hz. One octave down: $\frac{440}{2} = 220$ Hz.

The octave is a very big jump in pitch, so musicians break it up into 12 smaller steps. If the guitarist shortens the E string by one fret, the frequency will be $82.4 \times 1.059463 \approx 87.3$ Hertz.

Shortening the string one fret always increases the frequency by a factor of 1.059463. Why?

Because $1.059463^{12} = 2$. That is, if you take 12 of these hops, you end up an octave higher. This, the smallest hop in western music, is referred to as a *half step*.

Exercise 7 Notes and frequencies

Working Space

The note A near the middle of the piano, is 440Hz. The note E is 7 half steps above A. What is its frequency?

Answer on Page 68

3.2 Chords and harmonics

Of course, a guitarist seldom plays only one string at a time. Instead, they use the frets to pick a pitch for each string and strums all six strings.

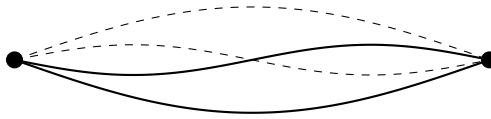
Some combinations of frequencies sound better than others, and the reason is simple whole-number ratios. We have already talked about the octave: If one string vibrates twice for each vibration of another, they sound sweet together.

Musicians speak of “the fifth”. If one string vibrates three times and the other vibrates twice in the same amount of time, they sound sweet together.

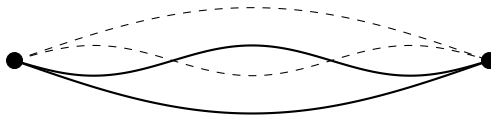
Likewise, if one string vibrates 4 times while the other vibrates 3 times, they sound sweet together. Musicians call this “the fourth.” And 5 vibrations against 4 make “the major third.” Play a major third and a fifth together over a root note (frequencies in the ratio 4 : 5 : 6), and you have a major chord, the happy-sounding chord at the heart of most pop songs.

Different frequencies bend different hair cells in the inner ear, so you can hear all six notes at the same time when the guitarist strums their guitar.

When a string vibrates, it does not create a single sine wave. Yes, the string vibrates from end-to-end, and this generates a sine wave at what we call *the fundamental frequency*. However, there are also “standing waves” on the string. One of these standing waves is still at the centerpoint of the string, but everything to the left of the centerpoint is going up, while everything to the right is going down. This creates *an overtone* that is twice the frequency of the fundamental.



The next overtone has two still points — it divides the string into three parts. The outer parts are up, while the inner part is down. Its frequency is three times the fundamental frequency.



And so on. 4 times the fundamental, 5 times the fundamental, etc.

In general, tones with many overtones tend to sound bright. Tones with just the fundamental sound thin.

Humans can generally hear frequencies from 20Hz to 20,000Hz (or 20kHz). Young people tend to be able to hear very high sounds better than older people.

Dogs can generally hear sounds from about 65 Hz to 45 kHz, which is why a dog whistle is silent to you but not to your dog.

Exercise 8 A perfect fifth

A string vibrates at 220 Hz. What frequency is a perfect fifth above it (ratio 3 : 2)? How does that compare with going up 7 half steps?

Working Space

Answer on Page 68

3.3 Making waves in Python

Let us make a sine wave and add some overtones to it. Create a file named `harmonics.py`.

```
1 import matplotlib.pyplot as plt
```

```
2 import math
3
4 # Constants: frequency and amplitude
5 fundamental_freq = 440.0 # A = 440 Hz
6 fundamental_amp = 2.0
7
8 # Up an octave
9 first_freq = fundamental_freq * 2.0 # Hz
10 first_amp = fundamental_amp * 0.5
11
12 # Up a fifth more
13 second_freq = fundamental_freq * 3.0 # Hz
14 second_amp = fundamental_amp * 0.4
15
16 # How much time to show
17 max_time = 0.0092 # seconds
18
19 # Calculate the values 100,000 times per second
20 time_step = 0.00001 # seconds
21
22 # Initialize
23 time = 0.0
24 times = []
25 totals = []
26 fundamentals = []
27 firsts = []
28 seconds = []
29
30 while time <= max_time:
31     # Store the time
32     times.append(time)
33
34     # Compute value each harmonic
35     fundamental = fundamental_amp * math.sin(2.0 * math.pi * fundamental_freq *
36         ↪ time)
37     first = first_amp * math.sin(2.0 * math.pi * first_freq * time)
38     second = second_amp * math.sin(2.0 * math.pi * second_freq * time)
39
40     # Sum them up
41     total = fundamental + first + second
42
43     # Store the values
44     fundamentals.append(fundamental)
45     firsts.append(first)
46     seconds.append(second)
47     totals.append(total)
48
49     # Increment time
50     time += time_step
51
52 # Plot the data
53 fig, ax = plt.subplots(2, 1)
```

```

54 # Show each component
55 ax[0].plot(times, fundamentals, label="440 Hz")
56 ax[0].plot(times, firsts, label="880 Hz")
57 ax[0].plot(times, seconds, label="1320 Hz")
58 ax[0].legend()
59
60 # Show the totals
61 ax[1].plot(times, totals)
62 ax[1].set_xlabel("Time (s)")
63
64 plt.show()

```

When you run it, you should see a plot of all three sine waves and another plot of their sum:

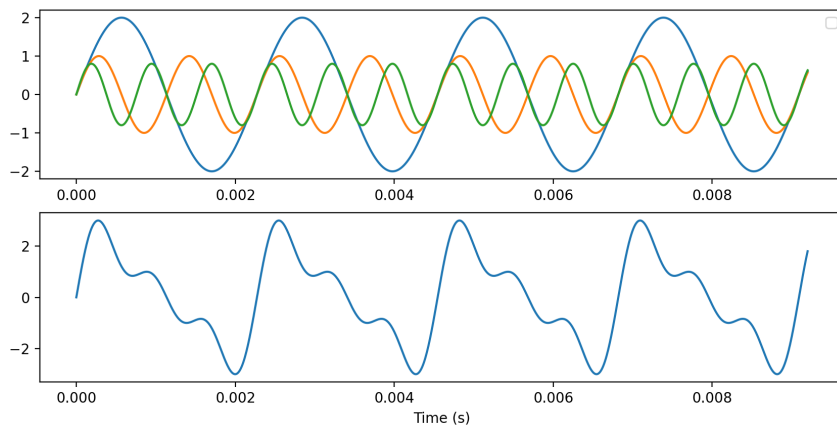


Figure 3.3: The output of harmonics.py

3.3.1 Making a sound file

The graph is pretty to look at, but let us make a file that we can listen to.

The WAV audio file format is supported on pretty much any device, and a library for writing WAV files comes with Python. Let us write some sine waves and some noise into a WAV file.

Create a file called `soundmaker.py`

```

1 import wave
2 import math
3 import random
4

```

```

5 # Constants
6 frame_rate = 16000 # samples per second
7 duration_per = 0.3 # seconds per sound
8 frequencies = [220, 440, 880, 392] # Hz
9 amplitudes = [20, 125]
10 baseline = 127 # Values will be between 0 and 255, so 127 is the baseline
11 samples_per = int(frame_rate * duration_per) # number of samples per sound
12
13 # Open a file
14 wave_writer = wave.open('sound.wav', 'wb')
15
16 # Not stereo, just one channel
17 wave_writer.setnchannels(1)
18
19 # 1 byte audio means everything is in the range 0 to 255
20 wave_writer.setsampwidth(1)
21
22 # Set the frame rate
23 wave_writer.setframerate(frame_rate)
24
25 # Loop over the amplitudes and frequencies
26 for amplitude in amplitudes:
27     for frequency in frequencies:
28         time = 0.0
29         # Write a sine wave
30         for sample in range(samples_per):
31             s = baseline + int(amplitude * math.sin(2.0 * math.pi * frequency *
32                 ↪ time))
33             wave_writer.writeframes(bytes([s]))
34             time += 1.0 / frame_rate
35
36         # Write some noise after each sine wave
37         for sample in range(samples_per):
38             s = baseline + random.randint(0, 15)
39             wave_writer.writeframes(bytes([s]))
40
41 # Close the file
42 wave_writer.close()

```

When you run it, it creates `sound.wav`, with four tones played quietly and then the same four played loudly. Each tone is followed by a burst of hiss. Open it in any audio player.

Notice how the tones are made: 16,000 times per second, the program computes the height of a sine wave and writes it as one number between 0 and 255. That is all digital audio is: a long list of numbers that tell a speaker where to be.

Exercise 9 One Octave Down

What is the frequency of the note one

Working Space

octave below 220 Hz?

Answer on Page 68

Scatter Plots and Correlation

Do students who sleep more get better grades? Do taller parents have taller kids? Does more experience mean a bigger paycheck? Each question is about two variables at once, which statisticians call *bivariate data*. In the last chapter, you summarized one variable with numbers like the mean and standard deviation. Now you will summarize the relationship between two.

4.1 Review of Scatter Plots

The first step with bivariate data is always a picture. In a *scatter plot*, each individual becomes one point: its x -coordinate is one variable and its y -coordinate is the other. When you look at a scatter plot, describe three things:

- **Direction:** Do the points trend upward (as x grows, y tends to grow) or downward?
- **Form:** Do the points follow a straight line, a curve, or no pattern at all?
- **Strength:** Do the points hug the pattern closely, or are they widely scattered?

Also watch for *outliers*, points that sit far from the overall pattern.

The variable you use to predict is the *explanatory* (or independent) variable, and it goes on the x -axis. The variable you are trying to predict is the *response* (or dependent) variable, and it goes on the y -axis.

4.2 Correlation Coefficient

A scatter plot lets you judge a relationship by eye. But how do we move beyond “this looks like a strong relationship” to something we can actually calculate and compare? That is exactly what the *correlation coefficient* gives us.

A correlation coefficient is a numerical measure used to judge the relation between two variables. The one we will use throughout this course is *Pearson’s correlation coefficient*, also known simply as the correlation coefficient. It measures the strength and direction of the linear relation between two quantitative variables.

You will encounter two versions of this value depending on whether you are working

with a full population or a sample drawn from one. When computed from a population, the correlation coefficient is written as ρ (read as “rho”). When computed from a sample, it is written as r . In this course, we will almost always be working with sample data, so r is the symbol you will use most.

No matter how strong or weak a relationship is, r always falls within a fixed range:

$$-1 \leq r \leq +1$$

Think about what those endpoints mean. A value of $r = +1$ or $r = -1$ means the points in a scatterplot fall perfectly on a straight line. A value of $r = 0$ means there is no linear relationship between the two variables at all. Every other value of r falls somewhere in between, and in the next subsection we will look at exactly how to interpret where r lands.

Exercise 10 Boundaries of r

Answer the following questions about the range of the correlation coefficient.

Working Space

1. A student computes $r = 1.3$ for a dataset. What does this tell you, without looking at the data?
2. Two researchers study the same population but one uses the full population and the other uses a random sample. What symbol does each researcher use for their correlation coefficient?
3. Can r equal exactly 0.5? Can it equal exactly -1 ? Explain what each of those values would mean about the data.

Answer on Page 68

Direction

Now that we know r always falls between -1 and $+1$, the first thing to read from it is direction. The sign of r tells you immediately whether the two variables move together or against each other.

A positive value of r indicates that as x increases, y also tends to increase. Taller fathers tend to have taller sons, so the scatterplot shows an upward trend and r is positive. A negative value of r indicates that as x increases, y tends to decrease. As a phone gets older, its battery life tends to decrease, so the scatterplot shows a downward trend and r is negative.

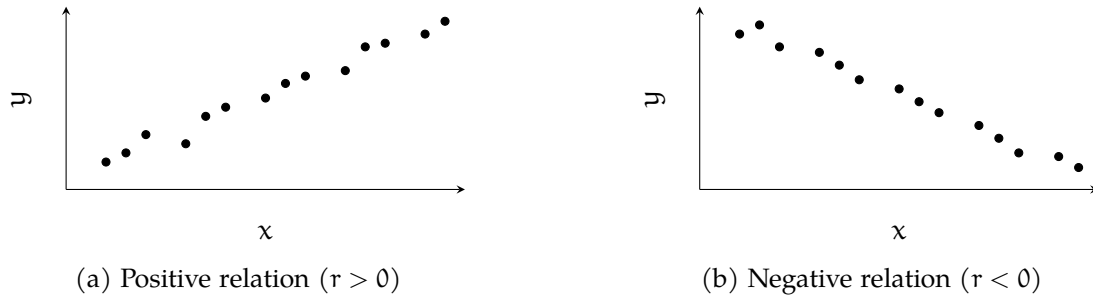


Figure 4.1: Direction of a linear relation as shown in a scatterplot.

Notice that direction says nothing about strength. Both $r = +0.1$ and $r = +0.95$ are positive, but they describe very different relationships. That is what we look at next.

Exercise 11 Reading Direction from r

For each pair of variables, predict whether r would be positive, negative, or zero.

Working Space

x	y	Sign of r
Hours studied	Exam score	
Temperature	Hot drinks sold	
Shoe size	Number of pets owned	
Age of a car	Resale value	

Answer on Page 69

Strength

Direction tells us which way a relationship goes, but not how strong it is. For that, we look at the numeric value of r , ignoring its sign. The further r is from zero, the stronger the linear relationship between the two variables. The closer r is to zero, the weaker it is.

At the extremes, $r = +1$ or $r = -1$ indicates a *perfect correlation*, meaning every point in

the scatterplot falls exactly on a straight line. In practice, real data almost never achieves this. On the other end, $r = 0$ means there is no linear relationship at all.

To compare the strength of two correlations, just compare the absolute values. If $r = -0.86$ for one pair of variables and $r = 0.75$ for another, the first pair has the stronger linear relationship because $0.86 > 0.75$.

Since interpreting strength can be subjective, statisticians often use cutoff values as a rough guide. The diagram below shows one commonly used set of cutoffs. Keep in mind that these are conventions, not rules. The key principle is always the same: the further from zero, the stronger the correlation.

Very Strong	Strong	Weak	No Correlation	Weak	Strong	Very Strong
-1.00 to -0.85	-0.85 to -0.50	-0.50 to -0.10	-0.10 to +0.10	+0.10 to +0.50	+0.50 to +0.85	+0.85 to +1.00

Figure 4.2: One commonly used set of cutoff values for interpreting the strength of a correlation coefficient.

Exercise 12 Interpreting Strength

The table below shows correlation coefficients computed from four different datasets. For each one, describe the direction and approximate strength of the

relationship.

r	Direction	Strength
+0.92		
-0.34		
+0.07		
-0.88		

Working Space

Answer on Page 69

Example

Let us put direction and strength together using a real dataset. Consider a sample of 10 father-son pairs. The heights (in inches) are recorded below.

Height of Father x (inches)	Height of Son y (inches)
66	66
66	65
67	67
67	70
68	67
68	70
69	70
70	72
71	72
73	74

Before computing anything, we check whether a correlation coefficient is appropriate. Correlation only applies to linear relationships, so we first make a scatterplot to confirm the relation between the two variables is linear.

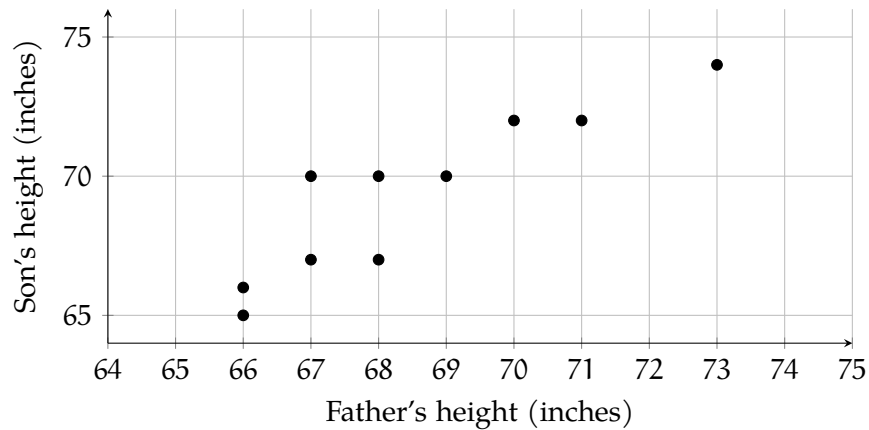


Figure 4.3: Scatterplot of father and son heights for a sample of 10 pairs.

The scatterplot shows a linear trend, so it is appropriate to compute r . In practice, correlation coefficients are computed using a calculator or statistical software rather than by hand. For this dataset, the computed value is approximately $r = 0.90$.

Computing r in Excel

Open your data in Excel or Google Sheets with the x values in column A and the y values in column B. To compute r directly, use the built-in correlation function:

1 `=CORREL(A2:A11, B2:B11)`

This returns the correlation coefficient between the two columns instantly. For the father-son data, this should return approximately 0.90.

Computing r in Python

```

1 import numpy as np
2
3 # Father and son heights in inches
4 x = [66, 66, 67, 67, 68, 68, 69, 70, 71, 73]
5 y = [66, 65, 67, 70, 67, 70, 70, 72, 72, 74]
6
7 r = np.corrcoef(x, y)[0, 1]
8 print(r)

```

`np.corrcoef()` returns a 2×2 matrix of correlation coefficients. The value at position `[0, 1]` is the correlation between `x` and `y`. For this dataset, the output should be approximately 0.90.

Try replacing the data with your own values and observe how r changes as the relationship between the two variables becomes stronger or weaker.

What does this tell us? The sign is positive, so taller fathers tend to have taller sons. The value is 0.90, which falls in the very strong range of our cutoff table. The points cluster tightly around a straight line, and the scatterplot confirms this.

Notice also what r does not tell us. It describes the strength and direction of the linear relationship, but not the shape of any curve, and not whether one variable causes the other. A strong correlation between two variables does not mean one causes the other to change.

Exercise 13 Interpreting a Computed r

Working Space

A researcher records the number of hours each of 40 students spent on social media in a week and their score on a science test. She finds $r = -0.62$.

1. Describe the direction and strength of the relationship.
2. A news story reports: "Social media lowers test scores!" Does $r = -0.62$ support that headline? Explain.
3. The researcher converts hours to minutes and recomputes r . What value

does she get?

Answer on Page 69

Correlation tells you how strongly two variables move together. The next chapter, The Regression Line, finds the line that best describes that relationship and uses it to make predictions.

The Regression Line

Do students who sleep more get better grades? Does a house with more square feet sell for more? In the Scatter Plots and Correlation chapter, you measured how strongly two variables move together with the correlation coefficient r . This chapter goes one step further: it finds the single straight line that best fits the data, so you can use one variable to predict the other, and it shows how to judge whether that line is any good.

5.1 Linear Regression

Suppose you wanted to predict a teacher's starting salary based on how many years of experience they have. You might expect that more experience leads to a higher salary, but how do you turn that intuition into a precise estimate? That is exactly what a *linear regression model* does. It gives you a straight-line equation that describes the relationship between two variables, so you can use the value of one to predict the value of the other.

The Regression Model

When one variable depends on or is explained by another, we call the variable being predicted the *dependent variable*. The variable doing the explaining is the *independent variable*. In our example, starting salary is the dependent (response) variable and years of experience is the independent (explanatory) variable. It would make little sense to say that a teacher's salary determines how many years they have worked.

The linear relation between two variables is described by the *population regression equation*:

$$Y = \alpha + \beta X$$

where Y is the response variable, X is the independent variable, α is the *y-intercept* (the value of Y when $X = 0$), and β is the *slope* (the amount Y changes for every one-unit increase in X). This equation describes the true relationship in the population, but in practice, we never know α and β exactly. We estimate them from sample data.

Once we have estimates for α and β , we can write the *estimated regression line*:

$$\hat{y} = a + bx$$

The symbol $\hat{y}$ (read “y-hat”) is the *predicted value* of Y for a given value of X . Notice the difference in notation: α and β are population parameters; a and b are their sample-based estimates.

Here is a quick reference for the symbols introduced so far in this chapter:

Quantity	Population symbol	Sample symbol
Correlation	ρ	r
Intercept	α	a
Slope	β	b

Real data never falls perfectly on a straight line. Two teachers with the same experience might earn different salaries, and neither will land exactly on the line. The difference between what the model predicts and what is actually observed is called the *residual*, or *error*, and is denoted e :

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

A positive residual means the observed value was higher than predicted; a negative residual means it was lower. From here on, salary will be expressed in thousands of dollars, so a value of 38 means \$38,000; this is the same convention the data table will use later in this section. Going back to our teachers: if the model predicts a starting salary of 38 for a teacher with four years of experience, but that teacher actually earns 40.5, the residual is $40.5 - 38 = 2.5$, or \$2,500. This teacher earned \$2,500 more than the model expected.

Exercise 14 Computing a Residual

A regression model predicts that a teacher with seven years of experience will have a starting salary of 42.3 (in thousands of dollars). The teacher’s actual starting salary is 40.1.

Working Space

1. Compute the residual.
2. Was the teacher’s salary over-predicted

or under-predicted? Explain.

Answer on Page 69

The goal of linear regression is to find the line that makes these residuals as small as possible across all observations.

The Least-Squares Regression Line

Every straight line you could draw through a scatterplot will produce a different set of residuals. Some lines will over-predict for certain points and under-predict for others. We want to know: which line does the best job overall? The answer is the *least-squares regression line*, also called the *line of best fit*. It is the line that minimizes the sum of the squared residuals, so, it makes $\sum e^2 = \sum (y - \hat{y})^2$ as small as possible.

Why should we square the residuals? First, squaring eliminates the sign, so positive and negative errors do not cancel each other out. Second, squaring penalizes large errors more heavily than small ones, which pushes the line toward a better overall fit. Now you know!

Exercise 15 Why Square the Residuals?

Suppose a regression line produces the following residuals for five data points: 3, -3, 2, -1, -1.

Working Space

1. Compute the sum of the residuals. What do you notice?
2. Compute the sum of the squared residuals.
3. Explain why simply minimizing the sum of the residuals would not be a useful goal.

Answer on Page 69

The least-squares regression line has two properties that always hold, regardless of the dataset.

First, it always passes through the point $(\bar{X}, \bar{Y})$ which represent the means (or average points) of the two variables. This point is sometimes called the *centroid* of the data.

Second, the slope b of the line is always given by:

$$b = r \left(\frac{S_y}{S_x} \right)$$

where r is the correlation coefficient, S_y is the standard deviation of the Y -values, and S_x is the standard deviation of the X -values. Notice that the slope depends directly on r : if there is no linear relationship ($r = 0$), the slope is zero, and the best prediction for any X is simply $\bar{Y}$.

Once you have b , the y -intercept a follows from the fact that the line passes through $(\bar{X}, \bar{Y})$:

$$a = \bar{Y} - b\bar{X}$$

These two formulas together give you the equation of the least-squares regression line:

$$\hat{y} = a + bx$$

Exercise 16 Computing the Line

For a dataset of teachers' years of experience (X) and starting salary in thousands of dollars (Y), the following summary statistics were computed: $\bar{X} = 6$, $\bar{Y} = 44$, $S_x = 3.2$, $S_y = 5.8$, $r = 0.91$.

Working Space

1. Compute the slope b of the least-squares regression line.
2. Compute the y -intercept a .
3. Write the equation of the least-squares regression line.
4. The line must pass through the centroid of the data. Verify this using

your equation.

Answer on Page 70

Interpreting Slope and Intercept

Having the equation of the least-squares regression line is only half the job. The other half is being able to explain what it means in plain language.

The slope b tells you how much the predicted value of Y changes for every one-unit increase in X . The key phrase you could see on an exam is **on average** because b is an estimate based on sample data, not an exact rule.

For our teachers example, if $b = 1.65$, the correct interpretation is:

For every additional year of experience, the predicted starting salary increases by \$1,650, on average.

Notice three things about that sentence. It names both variables. It gives the direction and amount of change. And it includes “on average.” Leave out any one of those and the interpretation is incomplete.

Exercise 17 Interpreting the Slope

A least-squares regression line for predicting a teacher’s starting salary (in thousands of dollars) from years of experience has slope $b = 1.49$.

Working Space

1. Write a correct interpretation of the slope in context.
2. A student writes: “For every extra year of experience, salary goes up by \$1,490.” What is wrong with this interpretation?

Answer on Page 70

The y-intercept a is the predicted value of Y when $X = 0$. Interpreting it requires care. Sometimes $X = 0$ is meaningful where a teacher with zero years of experience is simply a newly hired teacher. In that case the intercept has a natural interpretation: it is the predicted starting salary for someone entering the profession with no prior experience.

Other times $X = 0$ is outside the range of the data entirely, or physically impossible. In those situations, the intercept is mathematically necessary to define the line but has no useful real-world meaning. You should note that and move on.

Exercise 18 Interpreting the Intercept

Using the regression equation $\hat{y} = 34.10 + 1.649x$, where x is years of experience and $\hat{y}$ is predicted starting salary in thousands of dollars:

1. Write a correct interpretation of the y-intercept in context.
2. Is this interpretation meaningful here? Explain.

Working Space

Answer on Page 70

The Coefficient of Determination

Once you have a regression line, a natural question is how well it actually predicts Y . The *coefficient of determination*, written R^2 , answers this by measuring the proportion of the variation in Y that is explained by its linear relationship with X .

For a single predictor, R^2 is simply the square of the correlation coefficient:

$$R^2 = r^2$$

Because r always falls between -1 and $+1$, R^2 always falls between 0 and 1 . An R^2 close to 1 means the regression line accounts for nearly all of the variation in Y . An R^2 close to 0 means the line explains almost none of it.

R^2 is usually reported as a percentage. For example, a correlation of $r = 0.80$ gives $R^2 =$

0.64, so you would say that 64% of the variation in Y is explained by its linear relationship with X . The remaining $1 - R^2 = 36\%$ is left unexplained, due to other factors, natural variability, or measurement error.

Worked Example

A random sample of 10 teachers was selected from a large urban school district. Each teacher's years of experience at the time of hiring and their starting salary (in thousands of dollars) were recorded. The data is given in the table below.

Experience (years)	Starting Salary (\$1000s)
3	36
8	43
1	32
0	30
5	38
10	47
2	33
7	42
4	37
6	41

Before doing any calculations, look at the data. As years of experience increase, starting salaries appear to rise steadily. There are no obvious outliers, and the values are spread across a reasonable range of experience levels. A scatterplot (figure 5.1) confirms it: there is a strong positive linear relationship between experience and starting salary.

From the data, we have $n = 10$ pairs of measurements. Using a calculator or software, we can obtain the following summary statistics:

$$\bar{X} = 4.6, \quad \bar{Y} = 37.9, \quad S_x = 3.204, \quad S_y = 5.384, \quad r = 0.9957$$

Exercise 19 Slope, Intercept, and the Regression Equation

Using the summary statistics above:

Working Space

1. Compute the slope b of the least-squares regression line. Interpret it in context.
2. Compute the y -intercept a . Interpret

it in context.

3. Write the equation of the least-squares regression line.

Answer on Page 70

Exercise 20 Prediction, Residuals, and R^2

Using the regression equation $\hat{y} = 30.204 + 1.673x$:

Working Space

1. Predict the starting salary for a teacher with 5 years of experience.
2. The teacher in the sample with 8 years of experience has an actual starting salary of \$43,000. Compute the residual and interpret it.
3. Compute the coefficient of determination R^2 and interpret it in context.
4. About what percentage of the variation in starting salaries is *not* explained by years of experience? Suggest one factor that might account for this.

Answer on Page 71

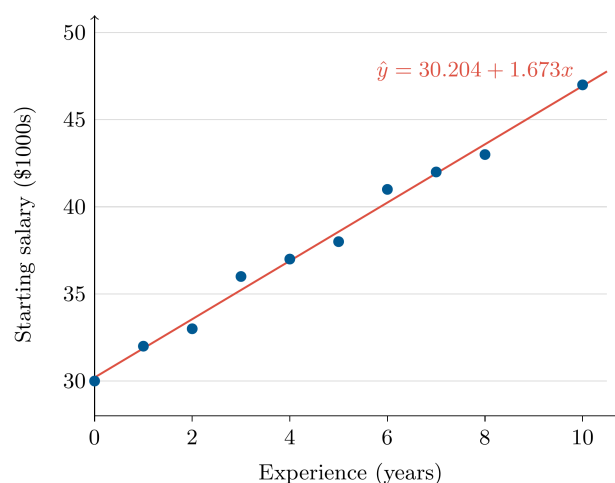


Figure 5.1: Experience and starting salary for the 10 sampled teachers, with the least-squares regression line $\hat{y} = 30.204 + 1.673x$.

Computing in Excel

You do not need to compute the slope, intercept, and R^2 by hand every time. Excel can produce all of these values at once, along with a scatterplot showing the line of best fit. Here is how.

First, enter your data into two columns, with experience values in column A and starting salary values in column B, with headers in row 1. Select both columns, then insert a scatterplot using **Insert** → **Chart** → **Scatter**. Once the chart appears, click on any data point, then select **Add Trendline**. In the trendline options panel on the right, choose **Linear** and check both **Display Equation on chart** and **Display R-squared value on chart**. Excel will draw the line of best fit and print the equation and R^2 directly on the chart.

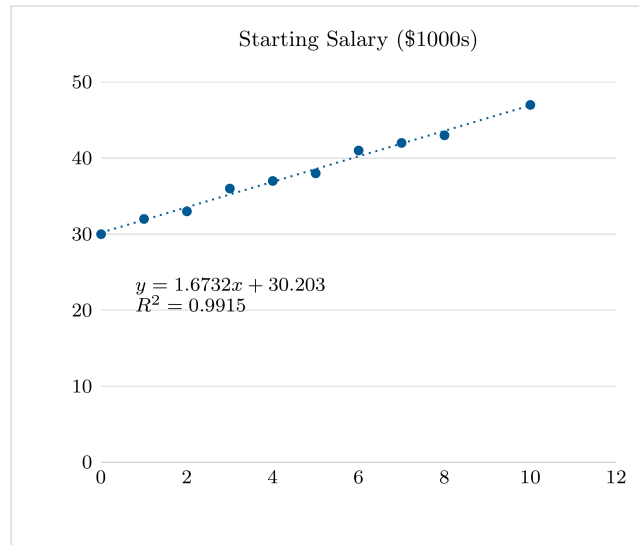


Figure 5.2: Excel scatterplot with a linear trendline, its equation, and R^2 displayed on the chart.

To get the full regression output, use the **Data Analysis** Toolpak. Go to **Data** → **Data Analysis** → **Regression**. Set the Input Y Range to your salary column and the Input X Range to your experience column. Check **Labels** if your columns have headers. Click OK and Excel will produce a full regression table in a new sheet. The values you need are in the **Coefficients** column: the row labelled **Intercept** gives you a , and the row labelled **Experience** gives you b . The R^2 value appears near the top of the output under **R Square**.

SUMMARY OUTPUT

Regression Statistics

Multiple R	0.9957
R Square	0.9915
Adjusted R Square	0.9904
Standard Error	0.5279
Observations	10

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	258.6706	258.6706	928.2004	1.46E-09
Residual	8	2.2294	0.2787		
Total	9	260.9000			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	30.2035	0.3028	99.7478	1.14E-13	29.5052	30.9017
Experience	1.6732	0.0549	30.4664	1.46E-09	1.5465	1.7998

Figure 5.3: Excel Data Analysis regression output for the teacher data. The highlighted cells give R^2 , the intercept a , and the slope b .

Computing in Python

Python can produce the same regression results as Excel, and a little more besides. The code below uses `scipy` for the regression, `numpy` to make the points of the line, and `matplotlib` to draw the scatterplot with the line of best fit.

```

1  import numpy as np
2  from scipy import stats
3  import matplotlib.pyplot as plt
4
5  # Data
6  experience = [3, 8, 1, 0, 5, 10, 2, 7, 4, 6]
7  salary      = [36, 43, 32, 30, 38, 47, 33, 42, 37, 41]
8
9  # Least-squares regression
10 slope, intercept, r, p, se = stats.linregress(experience, salary)
11
12 print(f"Slope (b):      {slope:.3f}")
13 print(f"Intercept (a): {intercept:.3f}")
14 print(f"r:              {r:.4f}")
15 print(f"R-squared:      {r**2:.4f}")
16
17 # Scatterplot with regression line
18 x_line = np.linspace(0, 10, 100)
19 y_line = intercept + slope * x_line
20
21 plt.scatter(experience, salary, color="steelblue", label="Data")
22 plt.plot(x_line, y_line, color="firebrick",
23         label=f"y = {intercept:.2f} + {slope:.2f}x")
24 plt.xlabel("Experience (years)")
25 plt.ylabel("Starting Salary ($1000s)")
26 plt.title("Experience vs. Starting Salary")
27 plt.legend()
28 plt.show()

```

Running this code produces the following output:

```

Slope (b):      1.673
Intercept (a):  30.203
r:              0.9957
R-squared:      0.9915

```

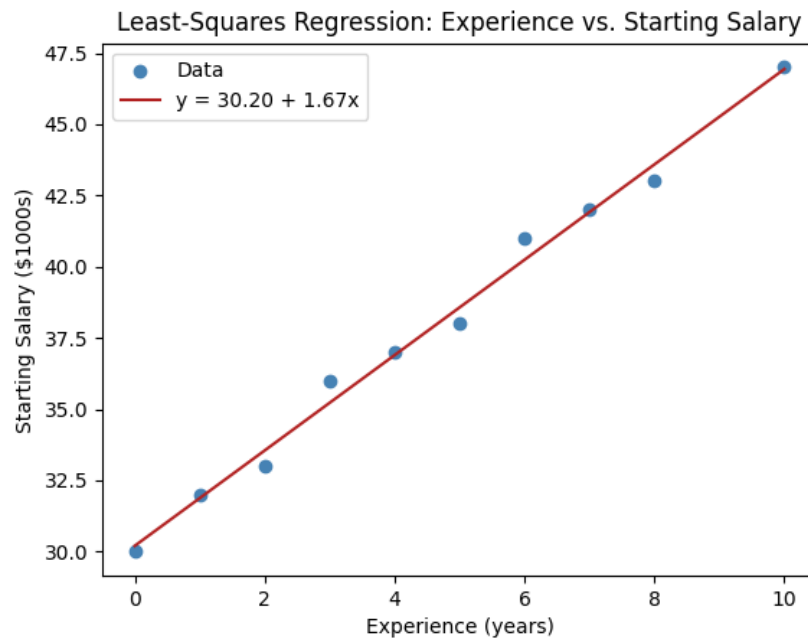


Figure 5.4: Experience vs. Starting Salary Graph generated from Python

These match the values we computed by hand, up to rounding. The `stats.linregress` function does all the heavy lifting! It returns the slope, intercept, correlation coefficient r , a p-value, and the standard error of the slope in one line. For now, focus on the first four outputs. The p-value and standard error will become important when we study inference for regression later!

Data Analysis with jamovi

A student club surveyed 100 classmates about sleep, screen time, breakfast, and a short quiz. The club wants to know what the answers show. This chapter introduces jamovi, a free program that does the statistics for you. You already know how to compute means, standard deviations, and regression lines by hand and in a spreadsheet; jamovi does them in a few clicks, and it can also run tests you will study in depth later.

6.1 Getting Started with jamovi

jamovi is a free statistics program. You choose an analysis from a menu, tick a few boxes, and the results appear right away.

You can use jamovi in two ways:

- In a web browser, at <https://cloud.jamovi.org>, with a free Guest account. Nothing needs to be installed.
- On your own computer, by downloading it from <https://www.jamovi.org>. This version works without the internet.

The jamovi window has three parts:

- On the left, the *spreadsheet* holds the data. Each row is one person or thing, and each column is one variable.
- Across the top, the *analysis ribbon* lists the analyses you can run.
- The *results panel* on the right shows tables and graphs.

To try it, open an example dataset. Click the menu icon (three lines) in the top-left corner, then choose Open, Data Library, Examples, and Tooth Growth. On the Analyses tab, click Exploration, then Descriptives, and move *len* into the Variables box. A table of statistics appears without pressing any Run button.

Exercise 21 Which Part of the Window?

Name the part of the window you would

Working Space

use for each task: spreadsheet, analysis ribbon, or results panel.

Task	Part
Fix a number that was typed wrong	
Choose an analysis	
Read the mean	

Answer on Page 71

6.2 Setting Up Your Data

Find `student_survey.csv` in your digital resources. Click the menu icon, choose Open, then Browse, and select the file. In jamovi Cloud, you may need to upload it first.

Column	What it records
<code>student_id</code>	A code number for each student
<code>travel</code>	How the student gets to school: Walk, Cycle, Bus, or Car
<code>breakfast</code>	Whether the student ate breakfast: Yes or No
<code>screen_time</code>	Hours on screens per day
<code>sleep_hours</code>	Hours of sleep per night
<code>sleep_quality</code>	The student's sleep rating: Poor, Fair, or Good
<code>quiz_score</code>	Score on a 20-point quiz

jamovi gives each column a *measure type*, which tells it what kind of data the column holds:

- Nominal: categories with no order, such as travel method
- Ordinal: categories with an order, such as Poor, Fair, Good
- Continuous: measured numbers, such as hours of sleep
- ID: labels for each case, such as a student code

jamovi guesses these types when it opens a file, and it can guess wrong. To fix one, double-click the column name and pick the correct type.

Two cells in `sleep_hours` are empty because those students skipped the question. jamovi treats empty cells as *missing values* and leaves them out of calculations.

Exercise 22 Measure Types

Give the correct measure type for each column.

Working Space

Column	Measure type
student_id	
breakfast	
sleep_hours	
sleep_quality	

Answer on Page 72

6.3 Describing Data

Click Exploration, then Descriptives. Move `quiz_score` and `sleep_hours` into the Variables box.

	quiz_score	sleep_hours
N	100	98
Missing	0	2
Mean	12.40	7.51
Median	12	7.60
Standard deviation	3.35	0.99
Minimum	4	4.5
Maximum	20	9.8

The average quiz score was 12.4 out of 20. Students slept 7.51 hours a night on average.

To compare groups, move a category into the Split by box. To draw a graph, open Plots and tick Histogram or Box plot. For a category such as `travel`, tick Frequency tables to count how many students are in each group.

Exercise 23 Breakfast and Quiz Scores

Put quiz_score in the Variables box and breakfast in the Split by box.

Working Space

1. What is the median quiz score for each group?
2. Which group did better?

Answer on Page 72

6.4 Comparing Two Groups

Students who ate breakfast averaged 13.28 on the quiz, and students who skipped it averaged 11.03. A t-test checks whether a gap this large could be due to chance.

Click T-Tests, then Independent Samples T-Test. Move quiz_score into Dependent Variables and breakfast into Grouping Variable. Under Tests, tick Welch's and untick Student's. Welch's version is the safer choice because it does not require the two groups to have the same spread.

jamovi reports $t = -3.40$ and $p = 0.001$. A small p-value, below 0.05, means the gap is unlikely to be due to chance alone.

The negative sign appears because jamovi subtracts the Yes group from the No group. The breakfast group still scored higher.

These data come from a survey, so they show a link between breakfast and quiz scores. They cannot show that breakfast causes higher scores.

Exercise 24 Reading the t-Test

Working Space

1. Is $p = 0.001$ below 0.05?
2. Write one sentence explaining what

the test shows.

Answer on Page 72

6.5 Looking at Relationships

Does more screen time go with less sleep? Click Exploration, then Scatterplot. Put `screen_time` on the X-Axis and `sleep_hours` on the Y-Axis.

The points slope downward: more screen time goes with less sleep.

Click Regression, then Correlation Matrix, and add both variables. The *correlation* r runs from -1 to 1 and measures how closely the points follow a straight line. Here $r = -0.70$, a strong negative relationship.

For a line of best fit, open Regression, then Linear Regression. Put `sleep_hours` in Dependent Variable and `screen_time` in Covariates. jamovi gives the line

$$\widehat{\text{sleep}} = 9.55 - 0.44 \times \text{screen time}$$

Each extra hour of screen time goes with about 0.44 fewer hours of sleep.

Exercise 25 Making a Prediction

Use the line to predict the sleep of a student with 6 hours of screen time.

Working Space

Answer on Page 72

6.6 Saving Your Work

Click the menu icon and choose Save As. jamovi saves an *.omv file*, which keeps your data and results together. In jamovi Cloud, download a copy to your own device before closing the tab.

To put a table or graph in a report, right-click it and choose Copy, then paste it into your

document.

Exercise 26 **Saving Choices**

Working Space

1. Which file type keeps both your data and your results?
2. How do you move a table into a report?

Answer on Page 72

With jamovi you can open data, set it up, describe it, test it, and save your results, all without writing code. Because it is free and runs in a browser, you can use it on almost any device.

Transforming Nonlinear Data with Logarithms

Certain relationships between two variables do not run in a straight line. Populations grow exponentially. Radioactive substances decay exponentially. When you try to fit a straight line to data like this, you may get a regression line, but the *line of best fit* will be wrong or non-existent. Logarithm models can straighten out curved relationships, allowing you to use linear regression techniques on data that is not linear in its raw form.

7.1 Exponential Models

A content creator posts a video and tracks how many views it gets each day:

Day (x)	0	1	2	3	4	5	6
Views, thousands (y)	4.2	8.3	17.4	34.1	77.8	161.8	337.5

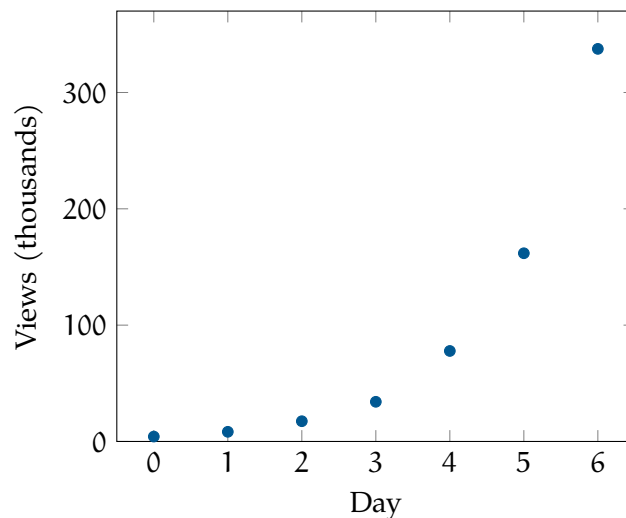


Figure 7.1: Daily views after posting. The curve bends upward instead of running straight.

A straight line fit to this data anyway would give $r^2 \approx 0.75$. That is not high enough to trust, and the residuals confirm it: the line misses low on both ends and high through the middle, a curved pattern rather than random scatter.

This would follow an exponential model, $y = a \cdot b^x$, where a is the starting count and b is the daily growth factor. Therefore:

$$\log(y) = \log(a \cdot b^x) = \log(a) + x \log(b)$$

$\log(a)$ and $\log(b)$ are both constants, so the right side is linear in x , with slope $\log(b)$ and intercept $\log(a)$. Graph $\log(y)$ against x instead of y against x , and a genuine exponential relationship becomes a line. This is a *logarithmic transformation*: the data does not change, only the scale you graph it on.

Apply that here:

Day (x)	0	1	2	3	4	5	6
Views, thousands (y)	4.2	8.3	17.4	34.1	77.8	161.8	337.5
$\log_{10}(y)$	0.623	0.919	1.241	1.533	1.891	2.209	2.528

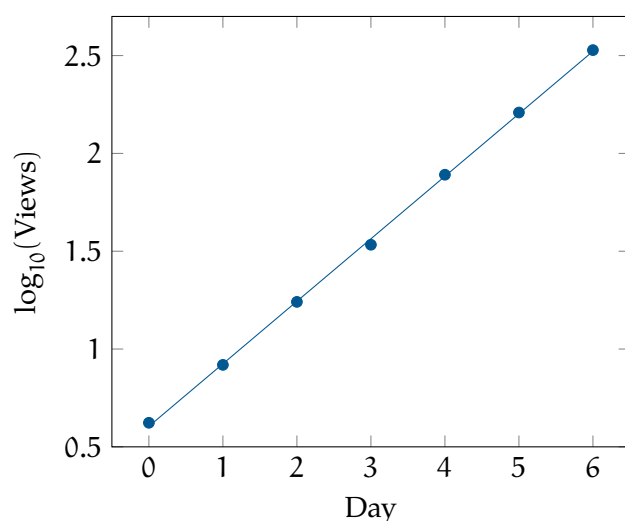


Figure 7.2: The same data after a logarithmic transformation. The points now fall along a line.

Regressing $\log_{10}(y)$ on x gives $\log_{10}(\hat{y}) = 0.3195x + 0.6050$, with $r = 0.9997$ and $r^2 = 0.9995$.

Slope = $\log_{10}(b) = 0.3195$, so $b = 10^{0.3195} \approx 2.087$. Intercept = $\log_{10}(a) = 0.6050$, so $a = 10^{0.6050} \approx 4.027$. That gives:

$$\hat{y} = 4.027 \cdot (2.087)^x$$

To predict, work in the transformed scale first, then convert back. At $x = 9$: $\log_{10}(\hat{y}) = 0.3195(9) + 0.6050 = 3.481$, so $\hat{y} \approx 10^{3.481} \approx 3,022$ thousand views.

- If $y = a \cdot b^x$, then $\log(y) = \log(a) + x \log(b)$, linear in x .
- Graphing $\log(y)$ against x turns a genuine exponential relationship into a line.
- The transformed slope equals $\log(b)$; the transformed intercept equals $\log(a)$.
- To predict y , find $\log(\hat{y})$ from the linear model first, then raise 10 to that power.

Computing in Python

Here is the video views example again, this time in Python:

```

1 import numpy as np
2 import statsmodels.api as sm
3
4 day = np.array([0, 1, 2, 3, 4, 5, 6])
5 views = np.array([4.2, 8.3, 17.4, 34.1, 77.8, 161.8, 337.5])
6
7 # Fit the raw model: views as a linear function of day
8 X = sm.add_constant(day)
9 raw_model = sm.OLS(views, X).fit()
10 print(raw_model.params)
11 print("R-squared:", raw_model.rsquared)
12 print(raw_model.resid)

```

This gives an intercept of -54.91 , a slope of 48.83 , and $R^2 \approx 0.750$.

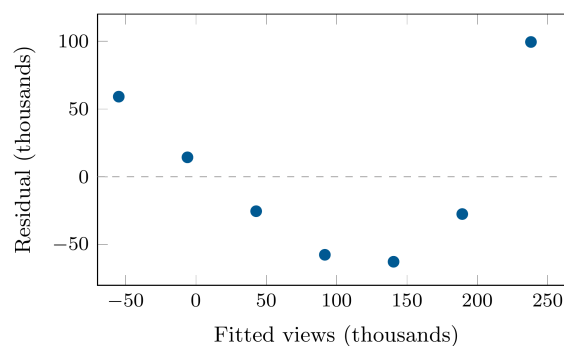


Figure 7.3: Residuals against fitted values for the raw linear model. The residuals form a U-shaped bow, so a line does not fit the data.

Apply the log transform and refit:

```

1 # Apply a natural log transform to the response variable
2 log_views = np.log(views)
3

```

```

4 log_model = sm.OLS(log_views, X).fit()
5 print(log_model.params)
6 print("R-squared:", log_model.rsquared)
7 print(log_model.resid)

```

Now the intercept is near 1.393, the slope near 0.7356, and $R^2 \approx 0.9995$.

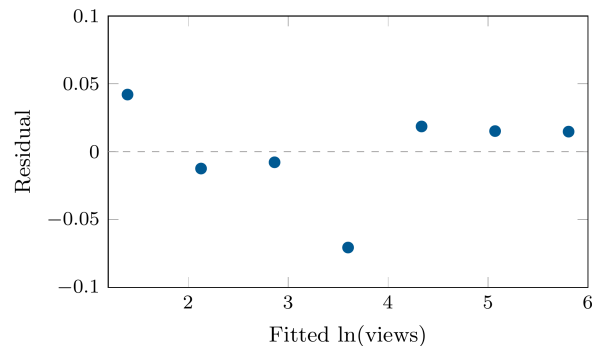


Figure 7.4: Residuals against fitted values after the log transformation. The residuals are small and show no curved pattern.

One detail changes between the by-hand work and the code: `np.log()` is the natural logarithm, not $\log_{10}$. Natural log pairs with `np.exp()` instead, so the transformed slope and intercept convert back a little differently:

```

1 slope = log_model.params[1]
2 intercept = log_model.params[0]
3
4 growth_factor = np.exp(slope)
5 percent_growth = (growth_factor - 1) * 100
6 starting_views = np.exp(intercept)
7
8 print(f"Growth factor per day: {growth_factor:.3f}")
9 print(f"Percent growth per day: {percent_growth:.1f}%")
10 print(f"Estimated starting views: {starting_views:.1f} thousand")

```

This gives a growth factor of about 2.087, a growth rate of 108.7 percent per day, and about 4.027 thousand starting views.

Exercise 27 Medication Concentration

Working Space

A patient takes a single dose of medication. A nurse measures the concentration remaining in the bloodstream every hour:

Hours (x)	Concentration, mg/L (y)
0	80.9
1	65.1
2	57.1
3	43.8
4	36.3
5	30.8
6	23.8

1. Find $\log_{10}(y)$ for each hour.
2. Regressing $\log_{10}(y)$ on x gives $\log_{10}(\hat{y}) = -0.0872x + 1.9108$, with $r = -0.9979$. Use this to write the exponential model $\hat{y} = a \cdot b^x$.
3. Predict the concentration at $x = 9$ hours. Is this interpolation or extrapolation?

Answer on Page 73

7.2 Power Models

Unlike exponential growth, certain curved relationships do not grow by a constant factor at every step. A marine biologist measures the length and weight of several fish from the same species:

Length, cm (x)	10	15	20	25	30	35	40
Weight, g (y)	15.7	53.9	135.5	244.0	461.4	668.0	1068.6

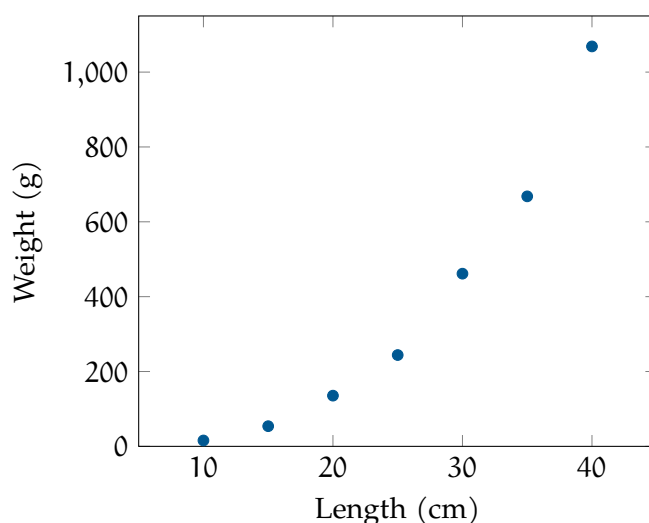


Figure 7.5: Fish weight against length. The curve bends upward, following a growth pattern distinct from exponential growth.

A straight line fit to this data yields $r^2 \approx 0.900$. Graphing $\log(y)$ against x , the same fix used in the last section, straightens the curve only slightly and leaves a pattern in the residuals. This relationship does not grow by a constant factor at every step the way the exponential model does; it grows by a constant power of x instead.

This kind of relationship follows a power model, $y = a \cdot x^b$. Take the logarithm of both sides:

$$\log(y) = \log(a \cdot x^b) = \log(a) + b \log(x)$$

Apply that to the fish data:

Length, cm (x)	10	15	20	25	30	35	40
Weight, g (y)	15.7	53.9	135.5	244.0	461.4	668.0	1068.6
$\log_{10}(x)$	1.000	1.176	1.301	1.398	1.477	1.544	1.602
$\log_{10}(y)$	1.196	1.732	2.132	2.387	2.664	2.825	3.029

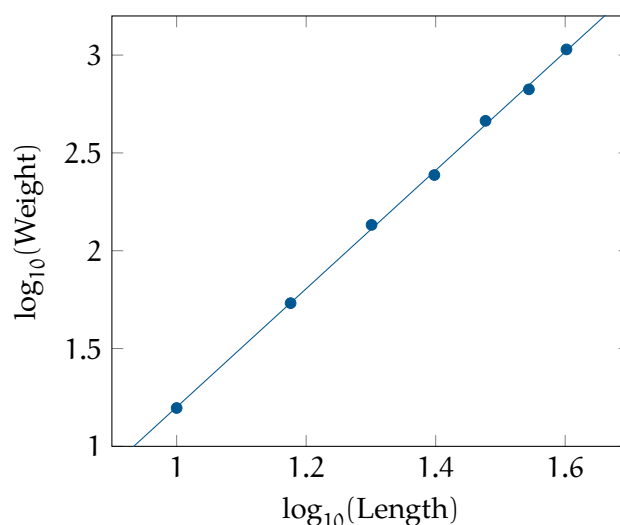


Figure 7.6: The fish data after a log-log transformation. Both axes are logarithms, and the points fall along a line.

Regressing $\log_{10}(y)$ on $\log_{10}(x)$ gives $\log_{10}(\hat{y}) = 3.0258 \log_{10}(x) - 1.8251$, with $r = 0.9997$ and $r^2 = 0.9993$, a much tighter fit than the raw linear attempt.

The slope is b itself, so $b \approx 3.026$. The intercept is $\log_{10}(a)$, so $a = 10^{-1.8251} \approx 0.01496$. That gives:

$$\hat{y} = 0.01496 \cdot x^{3.026}$$

An exponent close to 3 reflects real biology: fish grow in three dimensions at once, so weight scales close to the cube of length.

At $x = 45$ cm: $\log_{10}(\hat{y}) = 3.0258 \log_{10}(45) - 1.8251 = 3.177$, so $\hat{y} \approx 10^{3.177} \approx 1,503.9$ g.

- If $y = a \cdot x^b$, then $\log(y) = \log(a) + b \log(x)$, linear in $\log(x)$.
- Graphing $\log(y)$ against $\log(x)$, a *log-log plot*, turns a genuine power relationship into a line.
- The transformed slope equals b directly; the transformed intercept equals $\log(a)$.
- To predict y , find $\log(x)$ first, use the linear model to get $\log(\hat{y})$, then raise 10 to that power.

Exercise 28 Fall Time

Working Space

A student drops a ball from different heights and times how long it takes to hit the ground:

Height, m (x)	Fall time, s (y)
5	1.01
10	1.42
15	1.71
20	2.02
25	2.24
30	2.45
35	2.59

1. Explain why the shape of this scatter-plot points to a power model rather than an exponential one.
2. Regressing $\log_{10}(y)$ on $\log_{10}(x)$ gives $\log_{10}(\hat{y}) = 0.4908 \log_{10}(x) - 0.3388$, with $r = 0.9996$. Use this to write the power model $\hat{y} = a \cdot x^b$.
3. Predict the fall time from a height of 50 m.

Answer on Page 73

7.3 Choosing the Right Transformation

Both sections in this chapter start from the same move: something is curved, so take a logarithm and see if it straightens out. The two models call for a different logarithm, and mixing them up wastes time on a transformation that will not work.

Model	Equation	Transform	Linear in
Exponential	$y = a \cdot b^x$	$\log(y)$ vs. x	x
Power	$y = a \cdot x^b$	$\log(y)$ vs. $\log(x)$	$\log(x)$

When you are unsure which model fits, try both transformations and compare. Graph $\log(y)$ against x : if that looks straight, the relationship is exponential. If it still curves, graph $\log(y)$ against $\log(x)$ instead; if that straightens out, it is a power model. Whichever

transformation gives the higher r^2 with no leftover pattern in the residuals is the one that matches the data.

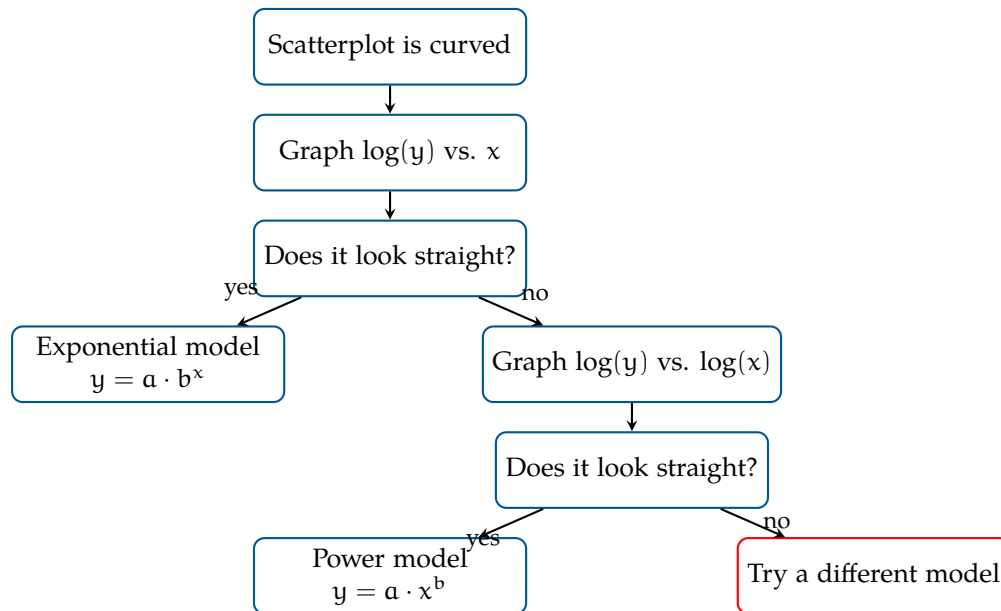


Figure 7.7: Deciding between an exponential and a power model.

These two transformations show up often outside the classroom: population growth tends to follow an exponential pattern, while relationships like body mass and metabolic rate, or a planet's distance from the sun and its orbital period, tend to follow a power model.

Answers to Exercises

Answer to Exercise 1 (on page 12)

$\frac{45}{0.1} = 450$ layers. $450 \times 30 = 13,500$ seconds, or 3.75 hours. Thinner layers look smoother but take longer.

Answer to Exercise 2 (on page 12)

Molding costs $20,000 + 0.5n$ and printing costs $8n$. They are equal when $20,000 = 7.5n$, or $n \approx 2,667$. Above about 2,667 stands, molding is cheaper. For 100 stands, print them (\$800 vs. \$20,050). For 100,000 stands, mold them (\$70,000 vs. \$800,000). This is why companies 3D-print prototypes and injection-mold products.

Answer to Exercise 3 (on page 17)

Amplitude: 15 cm (not 30; it is measured from equilibrium). Wavelength: 80 cm = 0.8 m. Frequency: $\frac{5}{2} = 2.5$ Hz. Period: $\frac{1}{2.5} = 0.4$ s. Speed: $(2.5)(0.8) = 2$ m/s.

Answer to Exercise 4 (on page 17)

$\lambda = \frac{v}{f} = \frac{343}{20} = 17.15$ m. That is longer than a school bus, which is why bass notes travel through walls so easily.

Answer to Exercise 5 (on page 19)

The speed is now $(5)(0.8) = 4$ m/s, twice as fast, so the crests move 0.4 m between curves (each 0.1 s apart), twice as far as before.

Answer to Exercise 6 (on page 19)

$\lambda = \frac{3 \times 10^8}{1 \times 10^8} = 3$ m. $T = \frac{1}{10^8} = 10^{-8}$ s, or 10 nanoseconds.

Answer to Exercise 7 (on page 24)

A is 440 Hz. Each half-step is a multiplication by $\sqrt[12]{2} = 1.059463094359295$. So the frequency of E is $(440)(2^{7/12}) \approx 659.26$ Hz.

Answer to Exercise 8 (on page 25)

A perfect fifth: $220 \times \frac{3}{2} = 330$ Hz. Seven half steps: $220 \times 2^{7/12} \approx 329.63$ Hz. They are very close but not identical. Modern instruments tune all twelve half steps equally (called *equal temperament*), so their fifths are very slightly flat.

Answer to Exercise 9 (on page 28)

Half of 220 Hz, which is 110 Hz.

Answer to Exercise 10 (on page 32)

1. A value of $r = 1.3$ is impossible. The correlation coefficient must satisfy $-1 \leq r \leq +1$, so the student made a calculation error.
2. The researcher using the full population uses ρ . The researcher using a sample uses r .
3. Both are possible. An r of 0.5 would indicate a moderate positive linear relationship. The points follow an upward trend but with noticeable scatter. An r of -1 would indicate a perfect negative linear relationship, meaning all points fall exactly on a downward-sloping line.

Answer to Exercise 11 (on page 33)

x	y	Sign of r
Hours studied	Exam score	Positive
Temperature	Hot drinks sold	Negative
Shoe size	Number of pets owned	Zero
Age of a car	Resale value	Negative

Answer to Exercise 12 (on page 34)

r	Direction	Strength
+0.92	Positive	Very strong
−0.34	Negative	Weak
+0.07	Positive	No correlation
−0.88	Negative	Very strong

Answer to Exercise 13 (on page 36)

1. Negative and strong: students who spent more time on social media tended to score lower.
2. No. Correlation does not show causation. Another variable could be behind both; for example, students with less time for homework might both use social media more and study less. Only a controlled experiment could show that social media itself lowers scores.
3. Still -0.62 . Changing units (multiplying every value by 60) does not change r , because correlation has no units.

Answer to Exercise 14 (on page 40)

1. $e = 40.1 - 42.3 = -2.2$, or $-\$2,200$.
2. The salary was over-predicted. The observed salary was $\$2,200$ less than the model expected.

Answer to Exercise 15 (on page 41)

1. $3 + (-3) + 2 + (-1) + (-1) = 0$. The positive and negative residuals cancel out, making the sum zero even though the line does not fit perfectly.
2. $3^2 + (-3)^2 + 2^2 + (-1)^2 + (-1)^2 = 9 + 9 + 4 + 1 + 1 = 24$.
3. A line that over-predicts some values and under-predicts others by equal amounts would have a residual sum of zero but it would be a poor fit. Squaring forces every error to contribute positively, so the minimization is meaningful!

Answer to Exercise 16 (on page 42)

1. $b = 0.91 \left(\frac{5.8}{3.2} \right) = 0.91 \times 1.8125 \approx 1.649$
2. $a = 44 - 1.649(6) = 44 - 9.897 \approx 34.103$
3. $\hat{y} = 34.103 + 1.649x$
4. Substituting $x = \bar{X} = 6$: $\hat{y} = 34.103 + 1.649(6) = 34.103 + 9.897 = 44.000 = \bar{Y}$. ✓

Answer to Exercise 17 (on page 43)

1. For every additional year of teaching experience, the predicted starting salary increases by \$1,490, on average.
2. The interpretation is missing the phrase “on average” and the word “predicted.” The slope is an estimate from sample data, not a guarantee that every teacher’s salary increases by exactly \$1,490 per year.

Answer to Exercise 18 (on page 44)

1. The predicted starting salary for a teacher with zero years of experience is \$34,100.
2. Yes, it is meaningful in this context. A teacher with no prior experience is a realistic case, so $X = 0$ is within the plausible range of the data.

Answer to Exercise 19 (on page 45)

1.

$$b = r \left(\frac{S_y}{S_x} \right) = 0.9957 \left(\frac{5.384}{3.204} \right) \approx 1.673$$

For every additional year of teaching experience, the predicted starting salary increases by \$1,673, on average.

2.

$$a = \bar{Y} - b\bar{X} = 37.9 - 1.673(4.6) \approx 30.204$$

The predicted starting salary for a teacher with no prior experience is \$30,204. Since $X = 0$ represents a newly hired teacher with no experience, this interpretation is meaningful in context.

3. The equation of the least-squares regression line is:

$$\hat{y} = 30.204 + 1.673x$$

where x is years of experience and $\hat{y}$ is the predicted starting salary in thousands of dollars.

Answer to Exercise 20 (on page 46)

1.

$$\hat{y} = 30.204 + 1.673(5) = 30.204 + 8.365 = 38.569$$

The predicted starting salary for a teacher with 5 years of experience is \$38,569.

2.

$$\hat{y} = 30.204 + 1.673(8) = 30.204 + 13.384 = 43.588$$

$$e = y - \hat{y} = 43 - 43.588 = -0.588$$

The residual is $-\$588$. The teacher's actual starting salary was \$588 less than the model predicted. The salary was slightly over-predicted.

3.

$$R^2 = r^2 = (0.9957)^2 \approx 0.9914$$

About 99.14% of the variation in starting salaries is explained by the linear relationship with years of experience. This is an exceptionally strong fit.

4. About $100 - 99.14 = 0.86\%$ of the variation is unexplained. This is very small, but it may be due to factors such as the teacher's level of education, the subject area taught, or the specific school within the district.

Answer to Exercise 21 (on page 51)

Task	Part
Fix a number that was typed wrong	Spreadsheet
Choose an analysis	Analysis ribbon
Read the mean	Results panel

Answer to Exercise 22 (on page 53)

Column	Measure type
student_id	ID
breakfast	Nominal
sleep_hours	Continuous
sleep_quality	Ordinal

Answer to Exercise 23 (on page 54)

1. Breakfast: 13 points. No breakfast: 11 points.
2. Students who ate breakfast had the higher median score.

Answer to Exercise 24 (on page 54)

1. Yes.
2. Students who ate breakfast scored higher on the quiz, and the difference is too large to blame on chance alone.

Answer to Exercise 25 (on page 55)

$9.55 - 0.44 \times 6 = 6.91$ hours of sleep.

Answer to Exercise 26 (on page 56)

1. An .omv file.

2. Right-click the table, choose Copy, and paste it into the report.

Answer to Exercise 27 (on page 60)

Hours (x)	0	1	2	3	4	5	6
$\log_{10}(y)$	1.908	1.814	1.757	1.641	1.560	1.489	1.377

Slope = $\log_{10}(b) = -0.0872$, so $b = 10^{-0.0872} \approx 0.818$. Intercept = $\log_{10}(a) = 1.9108$, so $a = 10^{1.9108} \approx 81.43$. The model is:

$$\hat{y} = 81.43 \cdot (0.818)^x$$

The concentration starts around 81.43 mg/L and multiplies by about 0.818 each hour, dropping roughly 18 percent per hour.

At $x = 9$: $\log_{10}(\hat{y}) = -0.0872(9) + 1.9108 = 1.126$, so $\hat{y} \approx 10^{1.126} \approx 13.37$ mg/L. The data only covers hours 0 through 6, so $x = 9$ is extrapolation.

Answer to Exercise 28 (on page 63)

The raw scatterplot rises quickly at first and then levels off, growing more slowly as height increases. Exponential growth does the opposite: it starts slow and accelerates. That mismatch in shape is the signal to try a power model instead.

Slope = $b \approx 0.4908$. Intercept = $\log_{10}(a) = -0.3388$, so $a = 10^{-0.3388} \approx 0.4584$. The model is:

$$\hat{y} = 0.4584 \cdot x^{0.4908}$$

An exponent near 0.5 matches the physics: fall time under gravity is proportional to the square root of height, $t \propto \sqrt{h}$.

At $x = 50$: $\log_{10}(50) \approx 1.699$, so $\log_{10}(\hat{y}) = 0.4908(1.699) - 0.3388 = 0.495$, giving $\hat{y} \approx 10^{0.495} \approx 3.13$ s.



INDEX

- y-intercept, 39
- .omv file, 55
- 3D printing, 10
- acoustics, 22
- additive, 10
- additive manufacturing, 10
- amplitude, 16
- an overtone, 24
- analysis ribbon, 51
- bent lamination, 8
- bivariate data, 31
- casting, 7
- centroid, 42
- coefficient of determination, 44
- compression, 18
- compressions, 18
- Computer Numerical Control, 5
- computer numerical control, 5
- correlation, 55
 - direction, 32
 - strength, 33
- correlation coefficient, 31
- crests, 16
- dependent variable, 39
- descriptive statistics, 53
- echo, 22
- equilibrium, 15
- error, 40
- estimated regression line, 39
- explanatory, 31
- flange, 7
- frequency, 16, 23
- fundamental frequency, 24
- hair cells, 22
- half step, 24
- hem, 7
- Hertz, 23
- independent variable, 39
- injection molding, 9
- jamovi, 51
- lathing, 5
- least-squares regression line, 41
- line of best fit, 41
- linear regression model, 39
- log-log plot, 63
- logarithmic transformation, 58
- longitudinal, 18
- longitudinal wave, 18
- malleability, 6
- measure type, 52
- milling, 3
- missing values, 53
- octave, 23

outliers, 31
overtone, 24

Pearson's correlation coefficient, 31
perfect correlation, 33
period, 16, 23
pitch, 23
population regression equation, 39
predicted value, 40

rarefaction, 18
rarefactions, 18
residual, 40
response, 31
results panel, 51

scatter plot, 31
slope, 39
speed of sound, 21
spreadsheet, 51
subtractive, 3
subtractive manufacturing, 3

t-test, 54
the fundamental frequency, 24
transverse, 18
transverse wave, 18
troughs, 16
turning, 5

wave, 15
wavelength, 16
Welch's t-test, 54

y-intercept, 39