CHAPTER 1

Modeling and simulation

1. Statistical models

A central concept to this course is that of a **model**. Broadly speaking, models are simplified representations of the world[1] There are many ways to represent models, but in science (and life), we often use **mathematical models**. For example, you might be familiar with Newton's equation:

$$F = ma \tag{1}$$

which related the force (F) acceleration (a) and mass (m) of a particle. This a mathematical model of the motion of a (non-relativistic) particle in a force field. While it isn't always true, it holds for such a wide range of applications that we call it a *fundamental* law of nature, and it turns out to be extremely powerful. We can combine Newton's equation with other fundamental lows build models of more complicated systems involving many particles. For example, to build a model of planetary motion, we can combine Newton's equation with his other law of gravity, which states that the gravitation force between two objects of masses m_1 and m_2 a distance r from each other is

$$F = G\frac{m_1 m_2}{r^2} \tag{2}$$

where G is a constant. We refer to these types of models – that is, those which are built upon fundamental laws of nature – as **mechanistic** models.

In statistics, we are often interested in problems where nothing remotely close to a fundamental laws exist (yet). Instead, they are based directly on observations (data) or our intuitions. We will call these **phenomenological models**. Such models still be quite useful provided we are aware of their limitations. Of course, there is not sharp distinction between mechanistic and phenomenological models, but the distinction is helpful nonetheless.

Often our models cannot make exact predictions about the value of a variable (this is true in both mechanistic and phenomenological models, but especially the latter). Instead, they only tell us the probability that a variable has a certain value, or falls within a range of values.

For example, if we were to construct a model of the (y) of the height a randomly selected pine tree in New Hampshire as a function of its age (x), we might begin by searching a relationship of the form

$$y = f(x) \tag{3}$$

Such a model might be relevant for conservation efforts, since it would be important to understand how trees develop over time and influence their surrounding environment. However, if we take a **sample** of the population, meaning we go out and measure the height of some trees for which we know the age, we will quickly find that trees of the same age can have different heights. This variation will be as result of many variables which are not in our model, e.g. the surrounding, the specific subspecies of pine, genetic variation within a subspecies to name just a few. In principle, we could construct a more complex model which includes, say, the DNA sequence of the tree (g):

$$y = f(x, g). \tag{4}$$

This model would presumably have less variation. That is, if we collect a sample of trees for which we know the height and DNA sequence, we would find that the variation between trees with the same DNA and height is less than the variation between trees with only the same height. Yet variation will remain unless we include all the variables which effect tree height. However, it is not so useful to include these variables, since we can't actually measure most of them and many will have only a small influence on the tree height. One approach to dealing with these "hidden" sources of variation is to define a **statistical model**, where y is not the same for each tree of the same height, but is instead a **random variable**. On type of a statistical model is a regression model

$$y = f(x) + \epsilon \tag{5}$$

where ϵ is a random variable (usually one that is zero, on average). Models of this sort are the topic of this course.

2. Random variables and distributions

The rigorous mathematical theory for random variables is very useful, but requires certain machinery which is beyond the scope of these notes. Fortunately, we go a long way without such formalism. For our purposes, a random variable can be understood as a variable which we cannot predict prior to an observation, regardless of how much information we have. We can define the space of **outcomes** as all the possible values that a random variable may take on. The outcomes for the roll of a dice are $1, 2, \ldots, 6$ for the dice, or positive numbers for the height of a tree. Usually the outcomes are numbers, even if we use a number to represent a non-numerical quantity (e.g.

[1] Model is sometimes used interchangeably with **theory**, although I usually think of models as being smaller in scope.

someone's gender). In probability theory, one distinguishes between outcomes and **events** – the latter are subsets of outcomes. For example, we might refer to the event that the roll of a die is grater than 2. It's good to be aware of these definitions, but you don't need to memorize them.

We can describe a characterize a random variable using a **probability distribution**, which maps a set of possible events to real numbers between 0 and 1. For example, suppose we ask a random student in the college whether they were born in the US. A probability distribution $P(Y)$ which *models* their answer is the **Bernoulli distribution**,

$$P(Y) = \begin{cases} q & Y = \text{YES} \\ 1-q & Y = \text{NO} \end{cases} \tag{6}$$

where q is the fraction of students in the college who were born in the US. More compactly, we can represent YES with 1 and NO with 0, and write

$$P(Y) = q^Y (1-q)^{1-Y}. \tag{7}$$

We say that q is a **parameter** in our model because regardless of it's value our model is still a Bernoulli distribution. It is very important that the sum of $P(Y)$ over all possible outcomes is 1 – this is simply saying that we are certain one of the outcomes will happen. We will use $P(1)$ to mean "the probability that $Y = 1$", or, in there is some ambiguity in which variable we are referring to, we might write $P(Y = 1)$.

In this context, q has a clear interpretation: it is the fraction of students in the college who were born in the US. The Bernoulli distribution is our default model for any variable that can take two possible outcomes, usually abstracted as 0 or 1. In order to state that a Bernoulli distribution is a model for some random variable Y, we write

$$Y \sim \text{Bernoulli}(q). \tag{8}$$

We might also say "Y *follows* as Bernoulli distribution" or "Y is a Bernoulli random variable". More generally, we say that a variable in a model follows a given distribution by writing

$$\text{Variable} \sim \text{Distribution}(\text{parameters}). \tag{9}$$

We will sometimes use θ to denote the parameters.

Turning back to the example of our survey, let's suppose we don't have information about every students in the college. Rather, a survey of five students from this class is conducted, finding 4 yeses and 1 no. What is our best prediction of the total fraction of students in the college who answered YES? What assumption do we make when we answer this question? The process of answering this question is **statistical inference**. More generally, we use statistical inference to make predictions about things we don't observe based one what we do observe (data).

3. Python as a tool for statistical modeling

When we generate samples using a computer we call them **simulations**. We will use python to perform simulations, and it is therefore important to have a basic understanding of the python language. It is assumed that you will go through the separate python tutorial notebook. For convenience, we will cover some basic tasks in this Notebook

Exercise 1: *Working with for loops*

3.1. Simulations. Here, we will focus on tools relevant for statistics. In Python, we can simulate random variables using the `numpy` library:

```
> import numpy as np
> q = 0.5
> y = np.random.choice(range(2),p=[q,1-q])
```

We can generate multiple samples using a for loop

```
> n_samples = 100
> y = np.zeros(n_samples) # makes an empty list (i.e. array) of n_samples zeros.
> for k in np.range(n_samples):
>   y[k] = np.random.choice(range(2),p=[q,1-q])
```

A simpler way of doing this is

```
> y = np.random.choice(range(2),n_samples,p=[q,1-q])
```

The more general form of this command is

```
> y = np.random.choice(range(k),n_samples,p=[q_1,q_2,...,q_k])
```

where $q_1 + \cdots q_k = 1$. This will generate a sample from

Exercise 2: *Building a probability model*

We can also generate simulations of more complex random variables using simple ones. In this case, it is useful to define a function in Python which generates samples of our new random variable. For, example:

Example 1. *Writing a function to run simulations of coin flips.*

Exercise 3: *Modifying existing code*

3.2. Visualization. An important tool for visualizing samples is a histogram. In python, we would write:

```
> plt.hist(samples,100,density=true)
```

The histogram shows us the frequency of different outcomes. Histograms are discussed here

3.3. Working with tabular data. Frequently, we will work with data in tabular form. We can do this using Numpy (hopefully you read about this in the python tutorial), e.g.

```
>  # imagine we have an array of times and corresponding temperature measurements:
> times = np.array([1,2,3,4,5])
> temps = np.array([72,71,75,75,73])
> # we can make a 2d numpy array
> data = np.transpose(np.array([times,temps]))
> data
```

The pandas package in python provides some additional functionality:

```
> # the pandas library provides a convenient way organize this data
> import pandas as pd
> df = pd.DataFrame(data,columns = ["time","tempature"])
```

Examples from class can be found here here.

3.4. Monte Carlo. Often, we run many simulations of a model in order to say something about the distribution without performing any analytical calculations. We call these **Monte Carlo** simulations. Monte Carlo simulations make use of the fact that we can always conceptualize probabilities as fraction of things. That is, if we have n samples of a variable Y and we want to estimate $P(Y = y)$, then we can count the number for which $Y = y$ – we denote this as $n(Y = y)$, and divide by the total number: $P(Y = y) \approx n(Y = y)/n$.

Example 2. *Running Monte Carlo simulations*

Questions concerning how many samples we need to generate to obtain meaningful estimates from Monte Carlo simulations will be addressed later on.

3.5. Means, variances, etc. There are ways in which we summarize attributes of random variables. If we have many samples $Y_1, Y_2, \ldots, Y_n$ of a random variable Y (e.g. answers to a survey question), the **sample mean** is defined as

$$\bar{Y} = \frac{1}{n}\sum_i Y_i \tag{10}$$

Often it is useful to quantify the deviations from the mean. Suppose each Y_i can take on outcomes $Y = 1, 2, 3, \ldots, m$. If n is large, then the fraction of samples for which $Y_1 = y$ will be $P(Y_1 = y)$, thus the sample mean converges to the true mean:

$$\bar{y} = \frac{1}{n}\sum_{y=1}^{m} y n_i = \sum_{y=1}^{m} y\frac{n_i}{n} \approx \sum_{y=1}^{m} yP(Y = y) \tag{11}$$

Sometimes we write $P(Y_i = y_i)$ and sometimes we write The expression on the right is the definition of the mean, or **expectation**, often denoted $\mathbb{E}[Y]$.

For this, we have the **sample variance**

$$\sigma^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2. \tag{12}$$

We will see why this makes sense as a measure of how spread our a distribution is later on when we talk about inference. For large n, this converges to the square root of the variance

$$\mathrm{Var}(Y) = \sum (y_i - \mathbb{E}[Y])^2 P(Y_i = y_i). \tag{13}$$

The sample standard deviation is:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_i (y_i - \bar{y})^2}. \tag{14}$$

In python, functions for implementing the mean and standard deviation are follows:

```
> np.mean(y)
> np.std(y)
```

By default, the standard deviation function divides the sum by the number of samples, we can fix this

```
> np.std(y,ddof=1)
```

Example 3. *Verifying an analytical formula with simulations*

Exercise 4: *Verifying an analytical formula with simulations*

We can always calculate the mean and standard deviation and mean of a sample regardless of the distribution it has been drawn from. However, we need to be careful, as the results may not be so meaningful. For example, the mean of a Bernoulli random variable is q, but (unless $q = 0$ or $q = 1$), the variable will never actually take on this value. For example, it might be more meaningful to think of the **mode**, which is the value that occurs most frequently.

4. Joint probabilities, Independence

We introduce, very briefly, the concepts of independence and conditioning. Just as we have considered single random variables, we can consider multiple random variables within the same model. Suppose we have two Bernoulli random variables Y_A and Y_B which model whether a person has mutations at two different locations on their genome. In this case, we need a model of both variables together:

$$\mathbb{P}(Y_A, Y_B) = \begin{cases} q_{00} & \text{if } Y_A = 0 \text{ and } Y_B = 0 \\ q_{01} & \text{if } Y_A = 0 \text{ and } Y_B = 1 \\ q_{10} & \text{if } Y_A = 1 \text{ and } Y_B = 0 \\ q_{11} & \text{if } Y_A = 1 \text{ and } Y_B = 1 \end{cases} \tag{15}$$

The probability distribution $P(Y_A, Y_B)$ tells us the probabilities for observing *both* variables together, e.g. observing a person with both mutations. It does not directly tell us the probabilities of observing e.g. someone with only one mutation. This can be obtained via marginalization; that is, summing over the other variable:

$$P(Y_A) = \sum_y P(Y_A, y) = P(Y_A, Y_B = 0) + P(Y_A, Y_A = 1) \tag{16}$$

where in the general the sum is taken over all possible outcomes for the second variable. $\mathbb{P}(Y_1)$ is defined similarly. For example,

$$P(Y_A = 1) = q_{10} + q_{11}. \tag{17}$$

This means that

$$Y_A \sim \text{Bernoulli}(q_{10} + q_{11}). \tag{18}$$

This is the distribution of Y_A absent any knowledge of Y_B. What if we are interested in the chance that someone has a mutation in gene A and we know they do not have a mutation in gene B? In this case, we introduce the **conditional probability** $P(Y_A = 1|Y_B = 0)$. This is defined as the chance that gene A has a mutation in a person if we know there is no mutation at gene B. If we want to think about this in terms of population averages, it is the fraction of mutations in gene A among only those people without mutations in gene gene B.

How do we calculate this? Using N to denote the number of individuals in a population with a given gene configuration,

$$P(Y_A = 1|Y_B = 0) = \frac{N(Y_A = 1, Y_A = 0)}{N(Y_B = 0)} = \frac{N(Y_A = 1, Y_A = 0)/n}{N(Y_B = 0)/n} \tag{19}$$

$$= \frac{P(Y_A = 1, Y_B = 0)}{P(Y_B = 0)} \tag{20}$$

This is a specific instance of Bayes' formula:

$$P(Y|X) = \frac{P(Y, X)}{P(X)} \tag{21}$$

Two variables are said to be **independent** if $P(Y|X) = P(Y)$.

Can you see why X being independent of Y implies Y is independent of X? Equation (**??**) is also true for events, for example, we will encounter things like

$$P(Y > z|X) = \frac{P(Y > z, X)}{P(X)} \tag{22}$$

For our purposes, it is important to understand the process of conditioning with data.

Example 4. *Showing independence*

Example 5. *Conditional averages*

Exercise 5: *Estimating conditional probability of dice*

Exercise 6: *Conditioning in gene model*

5. Additional exercises

Exercise 7: *Working with homocide data*

Exercise 8: *Simulating covid*

CHAPTER 2

Probability models

1. Binomial Distribution

A situation that often arrises is that we take many, say N, independent samples from a Bernoulli distribution. Now let Y be the number of 1s. Symbolically,

$$Y = \sum_{i=1}^{N} y_i, \quad y_i \sim \text{Bernoulli}(q). \tag{23}$$

Then Y follows **binomial distribution**:

$$Y \sim \text{Binomial}(N, q) \tag{24}$$

The binomial distribution has two parameters, N and p. Now let's think about the probability distribution. The chance to find any particular configuration of k ones is $q^k(1-q)^{N-k}$ because they are independent. For example

$$P(y_1 = 1, y_2 = 0, y_3 = 1) = P(y_1 = 1)P(y_2 = 0)P(y_3 = 1) \tag{25}$$

$$= q(1-q)q = q^2(1-q). \tag{26}$$

However, there are many configurations with k ones, in-fact there are

$$\binom{N}{k} = \frac{N!}{k!(N-k)!}, \tag{27}$$

and therefore

$$P(Y = k) = \binom{N}{k} q^k (1-q)^{N-k}. \tag{28}$$

The binomial distribution has a mean and variance

$$\mathbb{E}[Y] = qN \qquad \text{var}(Y) = Nq(1-q). \tag{29}$$

These formulas come from the fact that for sums of independent variables, the variance and expectation sum.

An important feature of the Bernoulli random variables is that the mean grows much faster in N than the standard deviation. This means that when N is very large, the deviations from the average will become very small relative to the mean. An important measure of variation relative to the mean is the coefficient of variation

$$\text{CV} = \frac{\sqrt{\text{var}(Y)}}{\mathbb{E}[Y]}. \tag{30}$$

Binomial samples can be generated in numpy with

```
> y = np.random.binomial(n,p,n_samples)
```

Often we are interested not in Y, but the fraction $\phi = Y/N$. For example, we might be interested in the vote share in an election. You should be able to see that $\mathbb{E}[\phi] = q$. What about the variance?

$$\text{var}(\phi) = \text{var}(Y/N) = \frac{1}{N^2}\text{var}(Y) = \frac{q(1-q)}{N} \tag{31}$$

Notice that this will tend towards zero as $N \to \infty$. Meanwhile, $\mathbb{E}[\phi]$ has no dependence on N. This is a very important property, as it allows us to determine q by approximating $\mathbb{E}[\phi]$ with the sample mean.

Example 6. *Coefficient of variation*

Exercise 9: *Generating binomial samples*

Exercise 10: *Binomial election modeling*

You should recognize that the assumption of independence is very important here. The following example illustrates an instance where this may break down for an election model. It is a bit contrived, but contrived examples, which we sometimes refer to as **toy models**, can be very helpful when it comes to building our intuition.

Exercise 11: *More election modeling*

2. Uniform distribution and probability density (optional)

A uniform random variable, denoted

$$Y \sim \text{Uniform}(a, b) \tag{32}$$

has an equal chance of taking any number in the interval $[a, b]$ (we assume $a < b$). Let $L = b-a$. This is distinct from other distributions we have encountered in that it is a **continuous distribution**, rather than discrete. For the uniform distribution,

$$P(y_1 \leq Y \leq y_2) = \frac{y_2 - y_1}{L} \tag{33}$$

for $a < y_1 < y_2 < b$. That is, the chance for Y to fall in any interval is simply the length of that interval. This insures that that the probability of Y being somewhere in $[a, b]$ is one: $P(a \leq Y \leq b) = 1$. Note that as $y_2 \to y_1$, $P(y_1 \leq Y \leq y_2) \to 0$. This tells us that the chance for Y to take any specific value is 0. Indeed, there are simply two many number (uncountably many) in any interval to assign positive probability to each. For continuous variables, it is sometimes useful to work with the density, $f(y)$ (we will use lower case letters for density and uppercase for probability distributions). $f(y)$ is the the probability per unit Y, meaning that if we look in a small interval

$$f(y)dy = P(y \leq Y \leq y + dy) = \frac{dy}{L}. \tag{34}$$

Thus, for uniform distribution the density is $1/L$ if $y \in [a, b]$ and 0 otherwise.

2.1. Joint density and conditional density. Conditioning works for probability density just as it does for probability distributions.

Example 7. *Conditioning with continuous variables*

2.2. Cumulative density function. Sometimes it is useful to characterize a continuous distribution not by the density, but by the **cumulative distribution function (CDF)**, defined as

$$F(y) = P(Y < y). \tag{35}$$

What is the CDF of the uniform distribution? The **median** is the value y_m for which $F(y_m) = 1/2$. What is the median of a Uniform distribution?

To better understand density and CDF, imagine a student says they will arrive at my office between noon and 3. Let Y represent the time a student arrives, which we will model as a Uniform random variable. Then the density is $f(y) = 1/3$ which has units 1/hours. We can think of f as the rate at which the CDF increases – that is, it is the velocity of probability.

3. Normal distribution and the central limit theorem

In the previous example, we say that if we take the average of many Bernoulli random variables, we get a histogram that looks a lot like a "bell curve" with a standard deviation was proportional to $1/\sqrt{n}$.It turns out this is true when we add up *any* sequence of independent and independent distributed random variables which are not too pathological (actually it is also true for many sequences of random variables which are not independent).

It is useful to define a special random variable which captures the statistical behavior of random sums. We call this a Normal random distribution

$$Y \sim \text{Normal}(\mu, \sigma). \tag{36}$$

We can generate Normal random variables in python with

```
> np.random.normal(0,1)
```

The Normal distribution is defined by the Gaussian

$$f(y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{-(y-\mu)^2}{2\sigma}}. \tag{37}$$

This is the classic bell curve shown in Figure **??**. The probability distribution for the Normal distribution is defined by the area under this curve. As I discuss in the previous section can think of f as the probability *per unit of the random variable*, e.g. probability/feet.

We use the curve above to calculate probabilities of events in the Normal distribution. For example

$$Y \sim \text{Normal}(5, 2) \tag{38}$$

what is (approximately) $P(Y > 7)$? Note that $5 + 2 = 7$, so this is asking how likely it is that a Normal variable is greater than 1 standard deviation above the mean. This about $13.5 + 2 = 15.5\%$

The **central limit theorem** tell us that when y_i are independent and have a finite mean and variance μ_y and σ_y and

$$\hat{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i, \tag{39}$$

then the distribution of $\hat{Y}$ should be close to that of

$$Y \sim \text{Normal}(\mu_y, \sigma_y/\sqrt{N}). \tag{40}$$

Example 8. *Comparing histograms*

Example 9. *Working with Normal random variables*

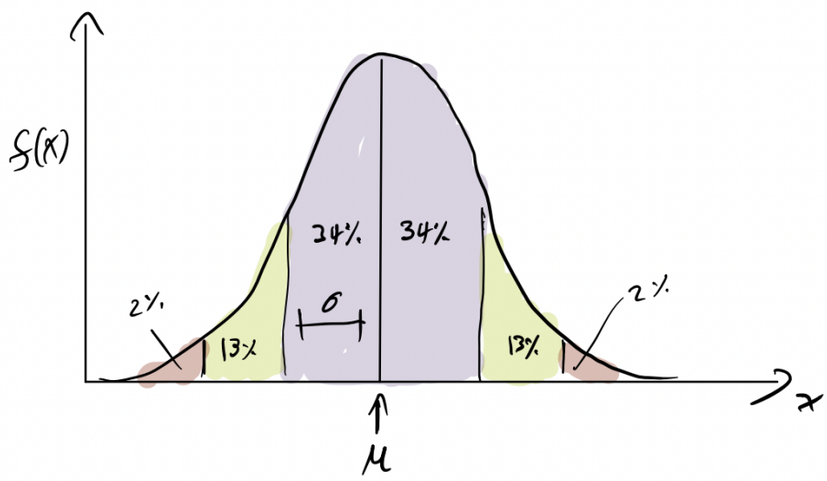

FIGURE 1. Probabilities in the Normal distribution

Exercise 12: *Hemoglobin levels*

4. Transformations of random variables

Now consider

$$X \sim \text{Normal}(\mu_x, \sigma_x) \tag{41}$$

and let

$$Y = aX + b \tag{42}$$

We are just multiplying and shifting everything. Think about what this does to the histogram (and try it in Python). Hopefully you can convince yourself that Y should also be Normal, but what are the mean and variance? Taking the average of both sides,

$$\mathbb{E}[Y] = a\mu + b \tag{43}$$

and

$$\text{var}(Y) = \text{var}(aX) + \text{var}(b) \tag{44}$$

Form the formula for variance, we know $\text{var}(aX) = a^2\text{var}(X)$. Also, $\text{var}(b) = 0$ So

$$Y \sim \text{Normal}(a\mu_x + b, |a|\sigma_x). \tag{45}$$

4.1. Standardizing.

We can transform any random variable into a so-called standard normal

$$Z \sim \text{Normal}(0, 1). \tag{46}$$

For defined above,

$$Z = \frac{X - \mu_x}{\sigma_x} \tag{47}$$

Then $a = 1/\sigma_x$ and $b = -\mu_x/\sigma_x$. Plugging into Equation (**??**) yields a standard Normal. **Transforming X to a standard Normal is equivalent to measuring X in units of standard deviations.** For example, if we make a histogram of X, all this transformation does is change the X axis to units of standard deviations from the mean.

5. Linear regression model

We now introduce the concept of regression modeling. A very broad class of models in statistics for the relationship between two variables X and Y is a regression model:

$$Y = f(X) + \epsilon \tag{48}$$

where f is a deterministic function; that is, if we evaluate f at a particular number, we get something that is not random. The term ϵ represents some source of noise **independent of** X, and is typically modeled with a Normal distribution

$$\epsilon \sim \text{Normal}(0, \sigma_\epsilon). \tag{49}$$

In other words, it represents things other than X which may influence Y. Regardless of how X is distributed, for any given values $X = x$, Y must have a Normal distribution:

$$Y|(X = x) \sim \text{Normal}(f(x), \sigma_\epsilon). \tag{50}$$

Of particular interest (due to its simplicity) is the case

$$f(x) = ax + b \tag{51}$$

which is the subject of this class. That is, we are interested in the model

$$Y = aX + b + \epsilon \tag{52}$$

where

$$\epsilon \sim \text{Normal}(0, \sigma_\epsilon). \tag{53}$$

5.1. Working with regression. In Equation (**??**), X could be anything, but let's suppose X is drawn from a Normal distribution. This gives us the model

$$Y = aX + b + \epsilon \tag{54}$$

$$X \sim \text{Normal}(\mu_x, \sigma_x) \tag{55}$$

(technically this is not a regression model anymore because we specify the distribution of X). Now that we have two random variables, we can ask questions like, what is the joint distribution? what are the conditional distributions? What are the Marginal distributions? Using properties of Normal random variables, we get

$$Y \sim \text{Normal}(a\mu + b, \sqrt{a^2\sigma_x^2 + \sigma_\epsilon^2}) \tag{56}$$

This is the distribution of Y regardless of X; that is, it is the distribution we would get if we randomly sampled Y values ignoring what the value of X is. What is another name for this? This distribution can be understood visually [DRAW DIAGRAM ON BOARD].

X and Y are not independent. This can be seen visually. **Note: even though X and Y are not independent, we don't say they effect each other. To see why this terminology is problematic note that Y has no "effect" on X (it is determined by X). However, what is $X|Y = y$?**

Example 10. *conditioning with continuous variables*

5.2. Interpretation of regression parameters. The interpretation of parameters in regression model is important. Sometimes (the example below) there is a clear meaning in terms of conditional distributions, but suppose we have a model of our time in a race as a function of the temperature:

$$Y = aX + b + \epsilon \tag{57}$$

Let's start with a: This is the average difference between times for races at temperature which differ by 1. (we expect this should be negative). ϵ is the variation around this average. Now what about b? The problem is that if we plug in $X = 0$ we obtain a nonsensical quantity, so really b does not have physical interpretation. This is common situation with regression models, and for this reason it is sometimes advantageous to center X.

Exercise 13: *Kid's test scores*

Exercise 14: *The random walk*

CHAPTER 3

Statistical inference

1. Estimators

The basic question of statistical inference can be framed as follow: We have a statistical model for a variable Y, e.g.

$$Y \sim \text{Normal}(\mu, \sigma), \tag{58}$$

but we don't know the parameters (in this case μ and σ). Why do care about the parameters? We need knowledge of these in order to make predictions.

Now imagine we have some samples of Y (either from simulations or data), $Y_1, Y_2, \ldots, Y_N$. **What are our best estimates of these parameters, and how accurate are they?** We've already tackled the first part of this problem for a number of distributions. The solution relies on two key observations:

(1) Both μ and σ can be represented as means over the distribution of Y. For example $\mu = \mathbb{E}[Y]$.
(2) If we have enough samples the sample average should be close to the actual average. That is, $1/N \sum_{i=1}^{N} Y_i \approx \mathbb{E}[Y]$. The central limit theorem tells us how accurate this estimate is.

To make this procedure more precise and generalizable, let's introduce some definition and notation. We will let $\hat{\theta}$ denote an **estimator** of a parameter θ from a sample if $\hat{\theta}$ is some function of our sample which is meant to approximate θ. For example

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} Y_i \tag{59}$$

is an estimator of μ in the Normal model. **Remember:** the estimator is a property of the data. That is, $\hat{\mu}$ **depends on the specific data we collect** or simulation we run. However, in classical statistics, it is meant to approximate something, μ which is a property of our statistical model. μ **does not depend on the data**.

Since $\hat{\mu}$ depends on the data, different replications of our sample will generate different values of $\hat{\mu}$. We can therefore think of $\hat{\mu}$ as a random variable. We call the distribution of $\hat{\mu}$ over many replications of our data the **sample distribution**. This is different than the distribution of Y, rather it is the distribution of $Y_1, Y_2, \ldots, Y_N$. For example, Y follows the Normal distribution given above, the sample distribution of $\hat{\mu}$ is

$$\hat{\mu} \sim \text{Normal}(\mu, \sigma/\sqrt{N}) \tag{60}$$

1.1. Standard errors.

A natural way to quantify the uncertainty in our estimate is the standard deviation of the $\hat{\theta}$ under the sample distribution. We call the resulting quantity the **standard error**, which is our **estimate** of the standard deviation of the sample distribution For the Normal model,

$$\text{se}(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{N}}. \tag{61}$$

This tells us how much our estimate will vary between different experiments (or surveys/simulations). The 95% **confidence interval**, or $95\% CI$ is the interval

$$[\hat{\theta} - 1.96\text{se}(\hat{\theta}), \hat{\theta} + 1.96\text{se}(\hat{\theta})] \tag{62}$$

In classical statistics, the interpretation of CI is subtle: it is not saying there is 95% chance or parameter will be in this interval. To understand why, note that the parameter has a 95% chance to be in the interval

$$[\theta - 1.96\text{std}(\theta), \theta + 1.96\text{std}(\theta)] \tag{63}$$

The standard errors and confidence intervals can help us decide how much data we need to collect.

Example 11. *Using standard errors to design an experiment*

There is an alternative way to think about the CI: as a measure of our belief in the parameter value. This interpretation is more natural for me, and we will discuss how to formalize it in the context of Bayesian statistics.

Exercise 15: *Standard errors for binomial model*

1.2. Bias and consistency. There must be some properties we would like the estimator to have. At a minimum, it should be in some way informed by the data. We express this with the assumptions that: The more data we have (e.g. the larger N) the closer we expect $\hat{\theta}$ to be to the true value. What do we mean by "closer" when we are talking about random things. This turns out to be technical, but for our purposes we will say $\hat{\theta}$ is **consistent** if

$$\mathbb{E}[\hat{\mu}] \to \mu \text{ and } \mathrm{se}(\hat{\theta}) \to 0 \text{ as } N \to \infty. \tag{64}$$

To better understand the notation of consistence, let's consider two rather silly ways to estimate q in a Bernoulli distribution. Let $\hat{q}_1$ and $\hat{q}_2$ be two other estimators of q defined by

$$\hat{q}_1 = \frac{Y}{N} + \frac{1}{N} \tag{65}$$

$$\hat{q}_2 = \frac{y_1 + y_2}{2} \tag{66}$$

Example 12. *Understanding consistency*

This example demonstrates that consistency is not the only property we look for in an estimator, since $\hat{q}_1$ seems inferior to $\hat{q} = Y/N$. To this end, we say that an estimator $\hat{\theta}$ is **unbiased** for some N (not just very large N), the average over the sample distribution is equal to the actual value under the model distribution; that is,

$$\mathbb{E}[\hat{\theta}] = \theta. \tag{67}$$

Exercise 16: *Understanding bias*

2. Maximum Likelihood

Sometimes it is quite clear what the estimator for a parameter should be, for example, this is the case for q in the Bernoulli distribution. However, we will find this is not always the case, so it is useful to have a **more systematic way of finding estimators.**

Recall that the probability distribution for the binomial distribution is

$$p(Y) = \binom{n}{Y} q^Y (1-q)^{n-Y} \tag{68}$$

In statistics, we sometimes call this the **likelihood**. More generally, the likelihood is defined as the probability we say a data set as a function of the parameters.

Equation (**??**) tells us how likely it is to observe k YES among n people surveyed. Then, it seems reasonable that this number should not be very small, since that would mean our survey results are an anomaly. More generally, the larger $\mathbb{P}(Y|q)$ is the more likelihood our results are. This suggests one a way to estimate determine q: We can take as our estimate $\hat{q}$ the value which makes $\mathbb{P}(Y|q)$ largest. In other words, we are finding the value of q which makes the data the most likely, and we will call this the **maximum likelihood estimate**.

You can do this using calculus (if you know how, I suggest you give it a try) to determine that the value of q which makes (**??**) largest is

$$\hat{q}_{\mathrm{MLE}} = \frac{Y}{n} \tag{69}$$

For a Normal distribution with mean and variance μ and σ, the MLE estimators are the usual sample mean and standard deviation which we have already been exposed to.

3. Estimators for parameters in a regression model

Consider the example of a clinical trial conducted as follows. Suppose N people participate in the trial, and are randomly assigned to the the control group (C) and treatment group (T) with probability $1/2$. People in T are given a drug whose effects is measure by a percent. We can model the distribution of blood pressure before and after treatment as

$$Y_C \sim \mathrm{Normal}(\mu_C, \sigma) \tag{70}$$

and

$$Y_T \sim \mathrm{Normal}(\mu_T, \sigma). \tag{71}$$

Our model for an individuals response can be framed as a regression model

$$Y = (\mu_T - \mu_C)X + \mu_C + \epsilon \tag{72}$$

where $X = 1$ if someone is in the treatment group and

$$\epsilon \sim \mathrm{Normal}(0, \sigma_\epsilon). \tag{73}$$

How would we estimate μ_T, μ_C and σ from a sample? Assuming we know the group each patient has been assigned to, we observe that

$$\mathbb{E}[Y|X=1] = \mu_T, \quad \sqrt{\mathrm{var}(Y|X=1)} = \sigma_\epsilon. \tag{74}$$

This means we can estimate these from sample averages of Y and X (and similar for $X = 0$). You might be tempted to say that the variance of Y is also σ_ϵ – but this is false! Why? (think back to the previous note when we looked at the marginal distribution of Y)

This means the sample distribution of the mean of the treatment group is

$$\hat{\mu}_T \sim \mathrm{Normal}(\mu_T, \sigma/\sqrt{N}) \tag{75}$$

If $\Delta\mu = \mu_T - \mu_C$, then as estimator $\Delta\mu$ is

$$\Delta\hat{\mu} = \hat{\mu}_T - \hat{\mu}_C \tag{76}$$

which is really the slope of the regression line. The sample distribution of $\Delta\hat{\mu}$ is also Normal:

$$\Delta\hat{\mu} \sim \text{Normal}(\Delta\mu, \sigma_\epsilon/\sqrt{N}) \tag{77}$$

3.1. Covariance and correlations coefficient. Now let's consider the general regression model

$$Y = aX + b + \epsilon \tag{78}$$

where X may be a continuous variable. **How would we estimate a?**

To understand how this can be done, we start by defining the covariance:

$$\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{79}$$

What is $\text{cov}(X, X)$?

$$\text{cov}(X, X) = \mathbb{E}[XX] - \mathbb{E}[X]\mathbb{E}[X] = \mathbb{E}\left[X^2 - \mathbb{E}[X]^2\right] = \text{var}(X). \tag{80}$$

If X and Y are independent, the covariance will be zero, it is possible for the covariance to be zero without the variables being independent.

Just like standard deviation and mean, estimates of covariance are obtained by replacing $\mathbb{E}$ with the sample average: That is, if we have samples $(x_1, y_1), (x_2, y_2), \ldots,$

$$\mathbb{E}[XY] \approx \frac{1}{N}\sum_i x_i y_i \tag{81}$$

In python, you can compute the covariance

```
> np.cov(x,y)[0,1]
```

The reason for the $[0, 1]$ is that the covariance function in numpy actually computes a 2D array (a Matrix), where the off diagonal entries are the covariance.

Example 13. *Covariance vs. independence*

3.2. Linear regression and least squares. For the linear regression model, we can show that

$$\text{cov}(X, Y) = a\sigma_x^2 \tag{82}$$

Since $\text{cov}(X, Y)$ and σ_x are both things that can be computed from a sample, this suggests an estimator for the slope variable

$$\hat{a} = \frac{\frac{1}{N-2}\sum_i (y_i - \hat{\mu}_y)(x_i - \hat{\mu}_x)}{\hat{\sigma}_x^2} \tag{83}$$

The $N - 2$ comes from the fact that we need at least two points to fit our regression model. We can rewrite this as

$$\hat{a} = \frac{\sum_i (y_i - \hat{\mu}_y)(x_i - \hat{\mu}_x)}{\sum_i (x_i - \hat{\mu}_x)^2} = \sum_i \left[\frac{(y_i - \hat{\mu}_y)}{(x_i - \hat{\mu}_x)}\right]^2 \frac{(x_i - \hat{\mu}_x)}{\sum_i (x_i - \hat{\mu}_x)^2} \tag{84}$$

This is weighted average of the rise/run between different points and the mean. Notice that if there are only two values of x, it reduces to the formula in the previous section. We can also show that

$$\hat{b} = \hat{\mu}_y - \hat{a}\hat{\mu}_x \tag{85}$$

$\hat{a}$ and $\hat{b}$ are also called the **least squares estimator**, because it happens to be the values of a and b which minimize the squared distance to the estimated regression line.

Since the samples are Normally distributed around the line with variance σ, we can estimate σ as the sample variance of the difference between our data and the line $\hat{b} + x\hat{a}$. However, since we are replacing the actual values of a and b with the estimations, we need to account for this in our estimation of the uncertainty and divide by $N - 2$ instead of N.

You never need to work with these formulas directly, as there is a Python package which does the computations for us. The code for fitting a linear regression model is

```
> X = sm.add_constant(x)
> model = sm.OLS(y,X)
> results = model.fit()
> print(results.summary())
```

The following example illustrates the use of this

Example 14. *Working with stats models*

Exercise 17: *More working with stats models*

CHAPTER 4

Working with regression models

1. Regression to the mean

Learning a little bit about the origin of the term regression can help us better understand regression models. Consider the regression model of daughter height (Y) conditioned on mother height (X):

$$Y \sim aX + b + \epsilon, \quad \epsilon \sim \text{Normal}(0, \sigma_\epsilon). \tag{86}$$

Over only one or two generations, we really don't expect the distribution of heights to change very much. Mathematically, this means Y and X should really have the same distribution. We call this the **steady-state** assumptions, because it amounts to the assumption that the distribution of heights is in a steady-state (which is a good approximation). Naively, we night expect that if the distribution of X and Y are the same, $a = 1$. This is because the steady-state assumption seems to suggest that the regression line should preserve the distribution, and therefore the average value $Y|X$ should be the same as the average of X. **This turns out to be false!**

To make sense of the fact that $a < 1$, let's do some math. We will suppose

$$X \sim \text{Normal}(\mu, \sigma) \tag{87}$$

The steady-state assumptions tells us:

$$Y \sim \text{Normal}(\mu, \sigma). \tag{88}$$

On the other hand, we have a formula for the standard deviation of Y (see week 2 notes):

$$Y \sim \text{Normal}(a\mu + b, \sqrt{|a|^2\sigma^2 + \sigma_\epsilon^2}). \tag{89}$$

The assumption that both distributions are equal implies

$$a\mu + b = \mu \tag{90}$$

$$|a|^2\sigma^2 + \sigma_\epsilon^2 = \sigma^2 \tag{91}$$

Solving these equations, we find that

$$b = \mu(1 - a) \tag{92}$$

$$\sigma = \frac{\sigma_\epsilon}{\sqrt{1 - a^2}}. \tag{93}$$

For this equation to make sense, $|a| < 1$, otherwise the standard deviation of the steady-state distribution explodes! The only exception is if $\sigma_\epsilon = 0$, since then y is a deterministic function of x.

What is the intuition behind all this? Let's imagine $a = 1$. Then abnormally tall mothers would birth to daughters that were on average just as tall (and the reverse for short mothers). This means that among all daughters, the spread of heights will be larger! The same thing will happen to the granddaughters and over time the standard deviation of the distribution of heights will continue to grow. We need $|a| < 1$ to balance out the effects of ϵ, which tends to spread things out. As a result, the average height conditioned on mother height is a combination of the mother's height and mean height among all mothers, μ:

$$\mathbb{E}[Y|X = x] = ay + (1 - a). \tag{94}$$

Example 15. *Simulation of an autoregressive process*

An important lesson from the autoregressive example is that **small differences is parameters can lead to HUGE differences in the results! It's crucial to understand what parameters**.

Exercise 18: *Working with autoregressive models*

2. Some basic model evaluation

Often we are interested in fitting data (i.e. inferring the parameters) to a linear regression model because we want to make predictions. How do we access how accurately we can make predictions? In order to address this questions it is very important we recognize there are different types of predictions we might want to make. For example, in the context of predicting the outcome of an election, we not interested so much in the distribution of outcomes, rather (since there is only one election). If we are designing a drug, it doesn't matter if there is an effect on the average if there is a very wide distribution of outcomes. We refer to predictions of SPECIFIC Y values as **point predictions**.

On the other hand, if we are interested in a scientific question, such as the heritability of human height, it is not so important whether we can predict individual heights, rather we are interested in understanding what the entire distributions of heights is. For example, we might want to know the chance that a person is greater than 6.5 feet. We will refer to these types of predictions – that is, predictions about the statistical behavior of a variable – as **probabilistic predictions**.

2.1. Coefficient of determination. Let's think about how we would evaluate our model's ability to make point predictions. Let's say we have fit a linear regression model and obtained $\hat{a}$, $\hat{b}$ and $\hat{\sigma}_\epsilon$. If we want to predict the value of Y given $X = x$, our best guess is

$$\hat{y} = \hat{a}x + \hat{b} \tag{95}$$

It is important to recognize that $\hat{y}$ depends on the data, just like $\hat{a}$ and $\hat{b}$. If we know the actual value of $Y|(X = x) = y(x)$, then we could look at the difference between the prediction and the actual value:

$$r = \hat{y} - y. \tag{96}$$

If course, we don't have y for every x only for the points in our data. Thus, a natural assessment of our models predictive power is to look at r_i for each data point:

$$S_r = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2. \tag{97}$$

S itself is not that useful though: it could be very large, and yet if it is much smaller than the overall variation in Y, we can still make accurate predictions.

$$S_y = \sum_{i=1}^{n} (y_i - \hat{\mu}_y)^2. \tag{98}$$

We compare there two quantities we obtain the **coefficient of determination**.

$$R^2 = 1 - S_r/S_y. \tag{99}$$

This is what statsmodels returns. **We can think of R^2 as a measure of how much variation in Y is explained by X.**

The coefficient of determination is actually related to a familiar quantity, the covariance. To see why, notice that if X follows a normal distribution, can rewrite it as

$$R^2 \approx 1 - \frac{\sigma_\epsilon^2}{\sigma_y^2} = \frac{\sigma_y^2 - \sigma_\epsilon^2}{\sigma_y^2} \tag{100}$$

$$= \frac{a^2\sigma_x^2 + \sigma_\epsilon^2 - \sigma_\epsilon^2}{\sigma_y^2} = \frac{a^2\sigma_x^2}{\sigma_y^2} = \rho^2 \tag{101}$$

Where $\rho = \mathrm{cov}(X, Y)/(\sigma_x\sigma_y)$ is known as the **correlation coefficient**. To understand why ρ is meaningful, notice that if the spread of X is very large relative to the spread in Y, a small value of a corresponds to a larger association between X and Y if we measure things in standard deviations.

Example 16. *Generating simulated data with different values of R^2*

We can better understand ρ in terms of the standardized variables. Let's assume that the X values in our regression model follow a Normal distribution and define the standardized variables

$$Z_y = \frac{Y - \mu_y}{\sigma_y}, \quad Z_x = \frac{X - \mu_x}{\sigma_x} \tag{102}$$

Here μ and σ are the marginal mean and standard deviation of the variable in the subscript. As you will show in the exercise below, ρ is the slope of the regression line of Z_y vs. Z_x. This helps us understand the meaning of ρ: it is there regression line we get if we translate our data to the origin and then rescale the axis to, roughly speaking, contain the bulk of our point cloud. Notice that $|\rho| < 1$ – why? This is related to regression to the mean: Both Z_x and Z_y have the same standard deviation.

Exercise 19: *Some calculations involving ρ*

Exercise 20: *Interpreting R^2 in the context of applications*

3. Visualizing uncertainty in regression models

Just like any parameters, there is some uncertainty in our estimates of parameters in a regression model. It is useful to visualize this when we plot the regression line, as is shown here.

4. Making decision with regression models

In statistics, we might infer parameters not because we are interested in specific values, but rather because we would like to use them to make a decision. For example, whether a candidate drug is worth moving to the next step in clinical trials. This problem is often framed in terms of **hypothesis testing**, in which we assign a probability to a particular hypothesis or its converse. The basic procedure of hypothesis testing is as follows:

(1) Compute something called a test statistic, $\hat{T}$, which like any estimator is simply some function of the data.
(2) Ask: how likely would we be to obtain a value AT LEAST at large at $\hat{T}$ IF our hypothesis was false. The result is the p-value.

Let's return the to the example of a clinical trial described in the previous weeks notes. For simplicity we will assume that $1/2$ there are N people in EACH group. Then

$$\hat{\mu}_C \sim \text{Normal}(\mu_C, \sigma/\sqrt{N}) \tag{103}$$

$$\hat{\mu}_T \sim \text{Normal}(\mu_T, \sigma/\sqrt{N}) \tag{104}$$

thus

$$\Delta\hat{\mu} \sim \text{Normal}(\Delta\mu, \sqrt{2}\sigma/N). \tag{105}$$

In this case, our null hypothesis will be that $\Delta\mu = 0$; that is, there is no effect of the drug. As our test statistic, we measure how far $\Delta\mu$ is from zero in standard deviations:

$$\hat{T} = \frac{\Delta\hat{\mu}}{\text{se}(\Delta\hat{\mu})} \tag{106}$$

Now, let $\Delta\mu_0$ be the random variable representing the effect under the null hypothesis. If we estimated $\Delta\mu$, we get the sample distribution

$$\Delta\hat{\mu}_0 \sim \text{Normal}(0, \sqrt{2}\sigma/N). \tag{107}$$

At this point we can answer the question posed in step 2. That is, we can answer the question: If the null hypothesis was true, how likely would we be to observe a value of $\hat{T}$ larger than the one we did. This defines the p-value:

$$p_v = P(\hat{T}_0 > |\hat{T}||\hat{T}) \tag{108}$$

where $\hat{T}_0$ is the test statistic corresponding to $\Delta\hat{\mu}_0$. **Note that the probability in the definition of p_v is taken over the distribution of $\Delta\hat{\mu}_0$, not $\hat{T}$. In this way, p_v, like $\hat{T}$ can be thought of as a random variable that depends on the data.**

If the p-value is very small, then it is highly unlikely we would have observed what we did when the null hypothesis was true. In this case, we can REJECT the null hypothesis as false. Usually some threshold is set for this, and if the p_v is below that threshold we say our result in statistically significant. On the other hand, **if p_v is not small, it does not necessarily mean the null hypothesis is true.**

Example 17. *p-values*

A result is said to be statistically significant if $p_v < 0.05$. Visually, we can see that $\Delta\mu$ is statistically significant exactly if 0 is not contained in the confidence interval!

4.1. Problems with p-values, hypothesis testing and statistical significance. Despite the widespread use of p-values, classical hypothesis testing and statistical significance, these concepts have some problems. This does not mean they are not useful, rather it is important to understand how they might be applied in appropriately in practice.

First, typically the null hypothesis is never true, that is it is never the case that two subpopulations are exactly equal – that is, that there is no effect. If we have enough data, we can almost always rule out the null hypothesis.

Exercise 21: *Behavior of p-values in N and effect size.*

A major issue in who statistical significance is used in practice, is that is can create a selection bias in the published literature, where effects sizes are almost always over estimates.

Exercise 22: *Bias in the literature*

Finally, a philosophical problem with statistical significance is that the difference between statistically significant.

Exercise 23: *Problems with statistical significance*

CHAPTER 5

Regression modeling with multiple predictors

1. Multiple predictor linear regression

The real power of regression comes when we work with models of the form

$$Y = b + \sum_{i=1}^{K} a_i X_i + \epsilon \tag{109}$$

$$\epsilon \sim \text{Normal}(0, \sigma_\epsilon) \tag{110}$$

where X_i is a set of K predictor variables. If we want to think about this in terms of conditional averages, then

$$Y|(X_1 = x_1, \ldots, X_K = x_K) \sim \text{Normal}\left(b + \sum_{i=1}^{K} a_i X_i, \sigma_\epsilon\right) \tag{111}$$

This is the simplest generalization of the single-predictor regression model to work with multiple predictors, although as we will see it is not the only generalization. We now want to answer all the questions we asked for the original regression model in the context of this model, such as:

(1) What assumptions are we making and how do we interpret the parameters a_i?
(2) What are estimators of the parameters from data?
(3) How accurate is our model at predicting new Y values based on X values?

1.0.1. *Multiple predictors in python.* Let's start by seeing how to work with multiple predictors in python The first step is to get the predictor variables in the correct format for statsmodels. Statsmodels wants us to input a multidimensional arrray

$$X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{bmatrix} \tag{112}$$

The ith column contains the predictors that go with our ith observation y. This will tell statsmodels to also include a constant term (the intercept) β_0 in our regression.

The following code will get our data in this format:

```
> X = sm.add_constant(np.transpose(np.array([x_hs,x_iq])))
```

Example 18. *Our first regression with multiple predictors*

2. Interpretation and estimation of the parameters

In order to interpret the parameters, it's easiest to work with just two predictors:

$$Y = b + a_1 X_1 + a_2 X_2 + \epsilon. \tag{113}$$

Let start by just looking at the deterministic equation:

$$y = b + a_1 x_1 + a_2 x_2 \tag{114}$$

This describes a flat surface in two dimensions as shown in Figure **??**

If we make a slide through the surface in the x_1 direction and look it at from the side, we see a line with slope a_1 (and similarly for x_2). Now back to the regression model. We can understand a_1 is the slope of Y vs. X_1 for fixed (conditioned on) X_2. **The fact that it doesn't matter which value of X_2 we condition is an assumption of the model**. Mathematically, we can write

$$a_1 = \mathbb{E}[Y|X_1 = (x+1), X_2] - \mathbb{E}[Y|X_1 = x, X_2]. \tag{115}$$

It is important that we condition on BOTH variables?

You might guess the coefficient a_1 is also $\text{cov}(Y, X_1)/\sigma^2_{x_1}$. After all, if we look a slice of the 2D planer function $y(x_1, x_2)$ along the x_1 direction, we get the same slope for all x_2. It stands to reason if we look at only the points in the x_1-y plane our regression slope would be a_1. **This argument assumes that when we change x_1, x_2 does not also change**. This is best understood with an example

Example 19. *Understanding the multiple predictors regression slopes*

The important thing is that when we increase x_1 we are ALSO increasing x_2.

If the usual relationship in terms of the covariance doesn't hold, is there a more general relationship expression for a_1 in terms of conditional averages. The answer is, of course, yes! To

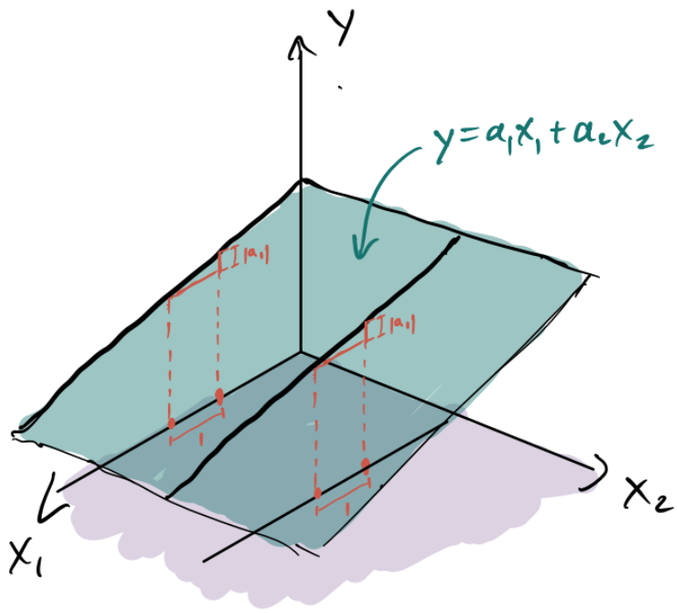

FIGURE 1. The function $y(x_1, x_2)$

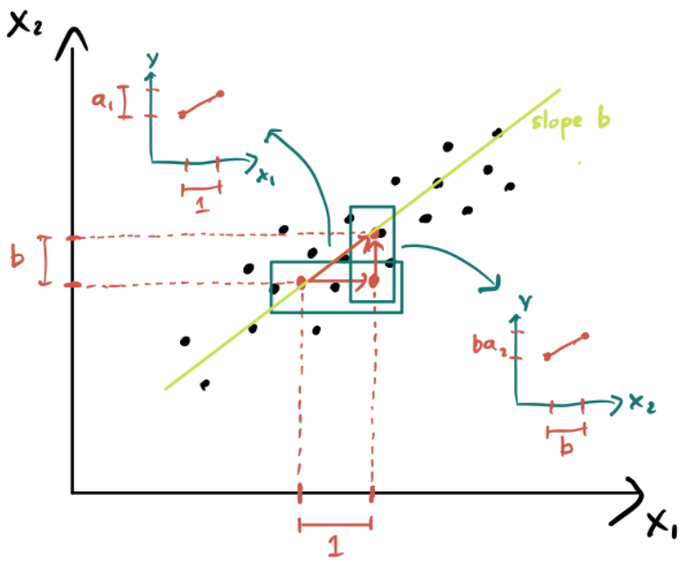

FIGURE 2. When we increase x_1 by 1, x_2 changes by b (which is the slope between x_1 and x_2 here, not the intercept.)

get there, we need to us some linear algebra which is beyond the scope of these notes. If you are interested, it goes something like this:

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sigma^2_{x_1} & \text{cov}(X_1, X_2) \\ \text{cov}(X_1, X_2) & \sigma^2_{x_2} \end{bmatrix}^{-1} \begin{bmatrix} \text{cov}(X_1, Y) \\ \text{cov}(X_2, Y) \end{bmatrix} \tag{116}$$

$$= \frac{1}{\sigma^2_{x_2}\sigma^2_{x_1} - \text{cov}(X_1, X_2)^2} \begin{bmatrix} \sigma^2_{x_2} & -\text{cov}(X_1, X_2) \\ -\text{cov}(X_1, X_2) & \sigma^2_{x_1} \end{bmatrix} \begin{bmatrix} \text{cov}(X_1, Y) \\ \text{cov}(X_2, Y) \end{bmatrix} \tag{117}$$

After using the formula for the inverse of 2×2 matrix, we obtain

$$a_1 = \frac{\text{cov}(X_1, Y)\sigma^2_{x_2} - \text{cov}(X_2, Y)\text{cov}(X_1, X_2)}{\sigma^2_{x_2}\sigma^2_{x_1} - \text{cov}(X_1, X_2)^2} \tag{118}$$

$$= \frac{\text{cov}(X_1, Y) - \text{cov}(X_2, Y)\text{cov}(X_1, X_2)/\sigma^2_{x_2}}{\sigma^2_{x_1} - \text{cov}(X_1, X_2)^2/\sigma^2_{x_2}} \tag{119}$$

You don't need to worry about this formula, but it essentially tells us how a_1 can be estimated from data: We replace the covariances and variances with the corresponding sample averages. Notice that if all the variances are equal to one:

$$a_1 = \frac{1}{1-\rho_{1,2}}(\rho_1 - \rho_{1,2}\rho_2) \tag{120}$$

where $\rho_{1,2}$ is the correlation coefficient between X_1 and X_2. Notice that if X_1 and X_2 are uncorrelated ($\rho_{1,2} = 0$), we obtain the usual connection between the regression coefficient and the correlation coefficient between X_1 and X_2.

Exercise 24: *Test score data*

Exercise 25: *More on test scores*

This can all be generalized to the situation where we have many predictors. The general formula for the regression coefficient would be:

$$\begin{aligned} a_i = \mathbb{E}[Y|X_1,\ldots,X_{i-1},X_i = x_i+1,X_{i+1},\ldots,X_K] \\ - \mathbb{E}[Y|X_1,\ldots,X_{i-1},X_i = x_i,X_{i+1},\ldots,X_K] \end{aligned} \tag{121}$$

We get a more complex expression for the coefficients but the idea is the same.

3. Collinearity and sloppy models

3.1. The sample distribution of coefficients. Just as before, we want to understand what the sample distribution of the coefficients looks like. In the multiple predictor case, this becomes more interesting, as the following example illustrates.

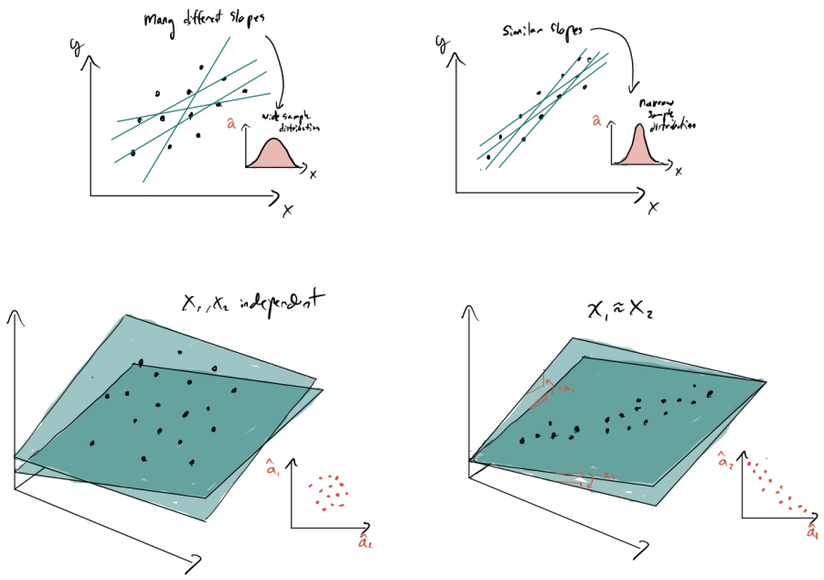

FIGURE 3. (top) In the single-predictor case, the width of the sample distribution measures how confident we are of a particular slope. It will be narrow if a replicate of our data is likely to produce a very similar slope. These means we get a rough idea of the width of sample distribution by seeing much we can change our regression line and still obtain something that appears to pass through our data. (bottom) In the two predictor case, we have a regression plane and changing a_1 and a_2 will "wiggle" the plane by tilting it in the x_1 and x_2 directions (there is also the intercept which can shift the plane up and down, but I'm not illustrating that). If X_1 and X_2 are uncorrelated, it doesn't matter which way we wiggle it, the fit will be similar, but if X_1 and X_2 are strongly correlated, wiggling the plane in the direction perpendicular to the points has a much smaller effect that parallel to them.

Example 20. *Understanding multivariate sample distribution*

To better understand what is going on, imagine X_1 and X_2 are very highly correlated (if they are perfectly correlated we say they are **colinear**). We can then write

$$Y = a_1X_1 + a_2X_2 + \epsilon \approx a_1X_1 + a_2X_1 + \epsilon \tag{122}$$

$$\approx (a_1 + a_2)X_1 + \epsilon \tag{123}$$

There are many ways to select a_1 and a_2 so that the surface $a_1x_1 + a_2x_2$ is close to the lines, since a change in a_1 can be compensated by a change in a_2. This means that **if we estimate** a_1 **and** a_2

and then generated new data, it would be possible to get a VERY different value of $\hat{a}_1$ and $\hat{a}_2$, so long as $\hat{a}_1 + \hat{a}_2$ is close to what we got before. This is illustrated in Figure **??** and Figure **??**. The following exercises explored in more depth what this means for the sample distribution.

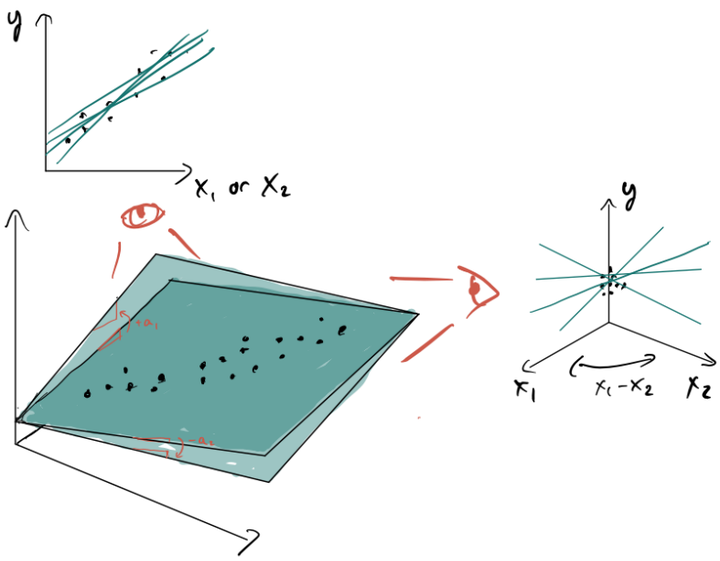

FIGURE 4. Different views of the data in the case when X_1 and X_2 are correlated. If we look at the data from the side, or along the $X_1 = X_2$ direction, then all our regression planes appear similar; however, when looked at from the "front" as shown in the right panel, we see that the places actually have very different slopes in the other direction.

Exercise 26: *Understanding multivariate sample distribution*

Exercise 27: *Sample distributions and predictors*

Exercise 28: *Implications for predictions*

3.2. Changing variables. At this point, you should understand that the sample distribution is related to correlations between x_1 and x_2. Indeed, for a large enough sample, one can show that

$$\hat{a}_1 \sim \text{Normal}\left(a_1, \sqrt{\frac{\sigma_\epsilon^2 \sigma_{x_1}^2}{\text{cov}(X_1, X_2)^2 - \sigma_{x_1}^2 \sigma_{x_2}^2}}\right) \tag{124}$$

Here, we can see explicitly what happens when X_1 and X_2 become highly correlated – the standard deviation of the sample distribution blows up. When this happens, we will say the model is **sloppy**. How do we deal with this situation? One approach is to use different predictor variables, for example, if $X_1 \approx X_2$, we might simply work with $X_1 + X_2$ as our predictor.

4. Dealing with categorical data

One situation in which models with multiple predictors frequently arrises is when trying to predict a Y variable based on categorical predictors, such as race. In this case, we need to transform the categories into numerical values. For example, if there are two cataogies, we map our variable to 0 or 1. If we have 3 categories, we might first think to map them to 0, 1 and 2. This has a problem though: A chance from 1 to 2 should not necessarily correspond to a change from 0 to 1. **There is no ordering of the x values**. Thus, instead we introduce a new variable, which is 1 if our data point is in the third category and zero otherwise. Do you see what the problem would be if we have 3 X variables, one for each category?

In order to take a categorical variable and transform it into a set of indicator variables in python, we use

```
> get_dummies
```

Example 21. *Working with categorical data*

Exercise 29: *Understanding marginal regression coefficients*

Exercise 30: *Simpson's paradox*

CHAPTER 6

Model assumptions

1. Model assumptions revisited

And this point you should understand

- The mathematical definition of a linear regression model, as well as the assumptions that are being made.
- How to fit regression models in python.
- How to interpret the results, including R^2, standard errors, confidence intervals and p-values.

This is only half of data science (or as I like to call it, science). The other half involves

- Building the “right" model to answer a given scientific question
- Integrating prior knowledge into the model and inference
- identifying deficiencies in our models
- and changing to them to answer the scientific questions we are interested in

1.1. General assumptions in statistical inference. Whenever we build a model and perform statistical inference, we are making assumptions about the the data. We've discussed a few of the assumptions we make in linear regression models, but in this section we are going to take a deeper dive into regression modeling assumptions. Some of them are explicit assumptions of the linear regression model, while others are more general assumptions we make in statistical analysis which are buried underneath the model itself, and often overlooked. We start by discussing these assumptions.

Validity: We assume that data is actually relevant to your research objective. For example, someones income does not necessarily tell you about someones total assets (they have a lot of debt, or simply be terrible at managing their money). So studying income can be misleading for certain research questions. This is often an issue when we study response variables that are aggregate statistics, such as metrics of performance. Do the aggregate statistics actually predict the results we are interested in? It's also an issue with subjective traits, like wellbeing, happiness. We might be able to find what factors are associated with someone reporting they are happy on a survey, but do these factors actually predict someone's long term happiness?

Representativeness: Whenever we fit a model on a finite data set and use it to make predictions about samples outsite the data set (e.g. future elections). We are assuming our sample is representative of the entire population (or at least the subset of the population we are interested in making predictions about). For example, if we fit a model using data from college basketball, will that same model be able to make predictions about the NBA? Maybe. If we can make predictions about elections in US, will we be able to predict the outcome of elections in the UK? Probably not.

1.2. Linear regression model assumptions. Below we will grow through our modeling assumptions in the linear regression context. It's important recognize that all these assumptions are **always false**. The question we must ask is whether they are adequate approximations for the questions we are interested in. If not, we need to understand how the data can be reorganized in a way that makes the linear regression model appropriate.

2. Normality of errors

We assume that the distribution of the errors is Normal. Why do we make this assumption? Partly for convenience as it's easy to work with Normal distributions, but on a deeper level, normal distributions emerge when noise is due to the additive contributions of many small sources of randomness. Mathematical, this is due to the central limit theorem. Roughly speaking, the Central limit theorem tells us that when noise is due to adding up many small source of randomness, we get a Normal distribution.

2.1. Missing a binary predictor. Consider for example adult male or female height h. Neglecting environmental factors, we can think of someone's height as being determined by their gnome:

$$h = \bar{h} + \sum_i \alpha_i g_i \tag{125}$$

where α_i is the effect of a mutation of the ith gene and g_i is 0 if a person has a mutation and 1 if not. There are tens of thousands of genes, so the central limit theorem tells us h has a Normal distribution. This is what we would call a toy model, meaning that it is not meant to actually describe the distribution of height quantitatively, rather, it serves to help us understand why something like height will have a Normal distribution.

On the other hand, if h is the height of **any** person, then

$$h = x_{\text{male}}\bar{h}_{\text{male}} + (1 - x_{\text{male}})\bar{h}_{\text{female}} + \sum_i \alpha_i g_i \tag{126}$$

Among *all* humans height will not have a Normal distribution because one particular factor, sex, has a very large effect.

So we can model height using a linear regression model with normal errors **as long as we include sex as a predictor.** In this case, our regression model is

$$h = ax_{\text{male}} + b + \epsilon \tag{127}$$

where

$$a = \bar{h}_{\text{male}} - \bar{h}_{\text{female}} \tag{128}$$

$$b = \bar{h}_{\text{female}} \tag{129}$$

$$\epsilon = \sum_i \alpha_i g_i \tag{130}$$

Normal noise also cannot model binary outcomes

If we want to model height as the response variable in a regression model we need to include sex as a predictor

2.2. Multiplicative randomness. Let's first work with a simple example involving one predictor: the effect of height on earnings.

Example 22. *Problems with earnings model*

Why is the assumptions of normality problematic for earnings? This is a situation where a "toy mode" can be very useful. A toy model for earnings is as follows. This model is not meant to have anything to do with the actual distribution of salaries, its only purpose is to illustrate a conceptual point that probably applies to the real distribution of salaries.

Let's imagine 1000 people **of the same height** enter the workforce at the same time each with a starting salary of y_0k. 20 years later there will of course be some variation in their earnings, represented by ϵ in the model. Let's think about what exactly causes that and the kind of distribution it might lead to.

A person might get lucky and gets a promotion, or is hired into a very prestigious position and so their salary will increase to say $2y_0$. Raises are generally some percent of a person's salary, so now if this person continues to be successful in their career, their salary will increase by an amount proportional to $2y_0$. After 20 years, the randomness in everyone's earning **will not simply be the sum of many small factors.**

To be mathematically precise, if someones gets a promotion that increases their salary by a factor ϕ_1 after their first year on the job, their salary the next year will be

$$y_1 = y_0 \times \phi_1 \tag{131}$$

If they get a promotion after their second year that increase their salary by a factor ϕ_2, their salary will be

$$y_2 = y_1 \times \phi_2 = y_0 \times \phi_1 \times \phi_2 \tag{132}$$

If someones get's 20 promotions over 20 years which increase their salary by factors $\phi_1, \phi_2, \ldots$ percent, their earning after 10 years will be:

$$y = y_0 \times \phi_1 \times \phi_2 \cdots \times \phi_{20} \tag{133}$$

Now let's simulate the salaries of 1000 people all get some sort of promotation or demotion each year. We will assume there is some variation in their promotions which we model by a Normal distribution:

$$\phi_i \sim \text{Normal}(1.01, 0.1) \tag{134}$$

This says that, on average, someone's salary goes up by 1%, but it could increase or decrease by as much as $\approx 20\%$.

Example 23. *Earnings toy model*

This example illustrates how non-normality can arrises from multiplicative effects. Recall that logarithms have the effect of transforming products into sums, thus:

$$\ln y = \ln y_0 + \sum_i \ln \phi_i \tag{135}$$

Intuitively, when we take the logarithm we are measuring our response variable in powers of e, or whatever the base of our log is (it doesn't matter).

2.3. Working on a log scale. Now back to our regression model. It makes more sense to model earnings on a log scale:

$$\ln Y = aX + b + \epsilon \tag{136}$$

where X is height. What a mean in terms of conditional averages? a is the average difference in **log earnings** between two people who differ in height by 1 inch. When we think about differences in the log of a response variable, we should remember that

$$\ln Y_1 - \ln Y_2 = \ln Y_1/Y_2. \tag{137}$$

so differences in log earnings correspond to log ratios between earnings.

We can also see this by exponentiating both sides of the linear regression equations. This yields

$$Y = e^{aX} e^b e^\epsilon \tag{138}$$

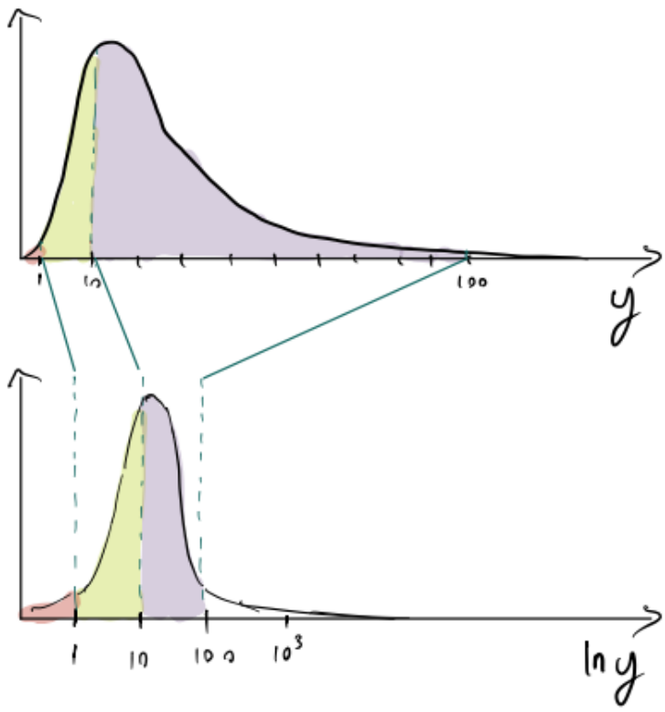

FIGURE 1. The effect of a log transform on the histogram for a very skewed distribution

The conditional average value of y is

$$\mathbb{E}[Y|X] = e^{aX} e^{b} \mathbb{E}[e^{\epsilon}] \tag{139}$$

That is, our model is saying that **if a person is one inch taller than someone else, they will make, on average,** e^a **times as much money**

If a is small (roughly between -0.4 and 0.4), then a useful approximation is

$$e^a \approx 1 + a. \tag{140}$$

Thus, **if a person is one inch taller than someone else, they make on average about** $100|a|\%$ **more (or less) money, assuming** a **is not too large**

Example 24. *Earnings*

Example 25. *Covid cases*

3. Independence of errors

In linear regression models, we generally assume the **errors** or noise values ϵ_i are independent. This assumption often fails when our predictor variables represent either time or space. The following example illustrates this.

Example 26. *Linear regression on unemployment data*

Exercise 31: *Simulating time series model for unemployment*

4. Residual plots

The previous examples illustrate how the residuals can be used to evaluate the linear regression model assumptions. Indeed, residual plots are central tool used to access modeling assumptions; however, there are some subtle aspects to their interpretation, particularly when working with multiple predictors. Let's take a mode systematic look at the use of residual plots.

The basic idea of residual plots is that by plotting the difference between the observed y values and the prediction of the $\mathbb{E}[Y|X]$, or

$$r_j = Y_j - \sum_i^K \hat{a}_i X_{i,j}, \tag{141}$$

we can identify any patterns that would suggest the assumption of the linear regression model are violated. In the instance of a single-predictor, we can simply plot r_j as a function of the predictor X. If we notice that the residuals do not appear to follow a normal distribution, or that the variance and mean change, then we should be skeptical.

When we have multiple predictors, what do we plot on the x axis? The answer is to plot r_j as a function of the predictors value of $\mathbb{E}[Y|X]$, or $\sum_i^K \hat{a}_i X_{i,j}$. The following example illustrates why.

Example 27. *Residual plots with multiple predictors*

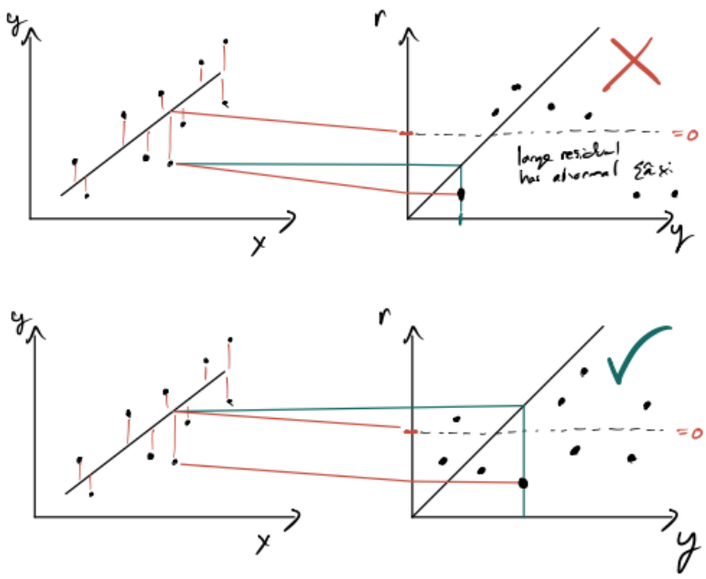

FIGURE 2. Correct and incorrect way to plot residuals against response variable

4.1. Identifying interactions in residual plots. In regression models, we assume that the response variable is, on average, a linear function of each of the predictors. Here we focus on "weak" nonlinearity, which comes from interactions between predictors rather than nonlinear dependence on the predictors themselves.

To understand what we mean by **interactions**, consider the model

$$Y = a_1 X_1 + a_2 X_2 + a_{1,2} X_1 X_2 + \epsilon \tag{142}$$

$$\epsilon \sim \text{Normal}(0, \sigma_\epsilon). \tag{143}$$

This is technically not a linear regression model in the predictors X_2 and X_2 because of the term involving $X_1 X_2$. The coefficient $a_{1,2}$ is called an interaction coefficient. How should we interpret this? First note that for fixed X_2 (respectively X_1), Y is a linear function of X_1 (respectively X_2). If we group the terms involving X_1, we can see what the Y vs. X_1 slope is for fixed X_2:

$$Y = (a_1 + a_{1,2} X_2) X_1 + a_2 X_2 + \epsilon \tag{144}$$

Thus

$$\tilde{a}_1(X_2) = a_1 + a_{1,2} X_2 \tag{145}$$

is a X_2 **dependent slope**. We can now see the meaning of $a_{1,2}$: is the average increase in the conditional slope of Y vs. X_1 corresponding an increase in X_2 by 1. We could have done the same calculation with X_1 and X_2 swapped though, so there is an equivalent interpretation in terms of the slope of Y vs. X_2.

Example 28. *Identifying an interaction*

Example 29. *Slopes in regression model with interaction terms?*

Exercise 32: *Interpreting residual plots*

CHAPTER 7

Model building

1. Linear models with features

The previous week was about assessing model assumptions and adjusting the model to correct for them. Here, we discuss how to build more complex models and directly access their predictive power on out-of-sample data (rather than indirectly access it by looking at model assumptions and R^2).

In the context of interactions, we already saw saw how a model can be extended by defining a new predictor $X_3 = X_1 X_2$. The more general idea that we can define a new predictor which is a function of the other predictors allows us to develop very complex and flexible models which nonetheless can be analyzed within linear regression framework. Here, we will formalize this, beginning with the case of a single predictor.

For a single prediction, consider the model

$$y = f(X) + \epsilon \tag{146}$$

A simple example would be $f(x) = b + ax + a_2 x^2$. This is simply a linear model if we define

$$X_1 = X, \quad X_2 = X_1^2. \tag{147}$$

More generally, a trick is to select a series of **basis** function, $\phi_1, \ldots, \phi_m$ and express f as a combination of them:

$$f(X) = \sum_{i=1}^{m} a_i \phi_i(X) \tag{148}$$

The function $\phi_i(x)$ are also called **features**.

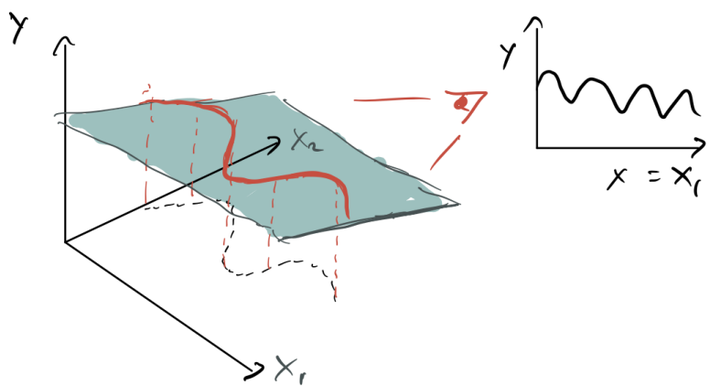

FIGURE 1. An illustration of how a nonlinear dependence on our predictor can be incorporated into the linear modeling framework by adding a feature.

Which functions ϕ should we use? The answer of course depends on the problem at hand. For example, we might know something about the physics of the data we are modeling. In some cases, we may select the ϕ so that the parameters a_i have clear interpretations (as is the case in linear regression model). The following illustrates such an example.

Example 30. *Mauna Kea Data*

Exercise 33: *Killing a tumor*

1.1. Orthogonality. It is often to our advantage to select basis function ϕ which contain very different "information". We know from before that ideally our predictor variables should be uncorrelated, otherwise $\hat{a}_i$ will be very correlated in the sample distribution. Thus, ideally, we should select ϕ_i so that the correlation coefficient between $\phi_i(X)$ and $\phi_j(X)$ is very small. When

$$\mathbb{E}[\phi_i(X)\phi_j(X)] = \mathbb{E}[\phi_i(X)]\mathbb{E}[\phi_j(X)] \tag{149}$$

we say that ϕ_i and ϕ_j are **orthogonal** with respect to the distribution of X. We won't get too deep into this, but it's important to understand when you go looking for basis functions. One can show for example, that the basis function $\phi_i(x) = \sin(2\pi x)$ are orthogonal with respect to

$$X \sim \text{Uniform}(-1, 1). \tag{150}$$

I won't say much more on this.

2. Overfitting

2.1. Direct assessment of model predictions. We have seen that any model has deficiencies and that **expanding our model always increases** R^2. A natural question is: Why not make our model as complex as possible? That is, why not add as many variables as we can and nonlinear terms? A simple answer could be given with the example of a linear regression: We can't draw a line through only two points. Similarly, we can't find a unique plane in d dimensions that goes through $< d$ points. This suggests we must not have more parameters in our model than data. Moreover, as the number of parameters in our model approach the amount of data, we become unable to resolve the parameters (the sample distributions become too wide) and interpret them in a meaningful way.

However, if our goal is make predictions with a model, there is much more to the story. Even before we have nearly as many parameters as data points, our model looses its power. This is not something we can see based on the behavior of error between the model and the data, such as R^2, which always decreases as we expand our model. Instead, we need to look at a **direct assessment of out-of-sample predictive power**. Ideally, we could look at the error between the model and the predictors for new X values. One way to access this is to break our data up into two subsets, a **training** (denoted $Y_{\text{train},i}$) and **test** set (denoted $Y_{\text{test},i}$) . We fit the model using only the training set, and then see how well our model can predict the values in the test set. The following examples reveals what we can learn from this:

Example 31. *Fitting polynomial data*

2.2. Bais-variance tradeoff. To make sense of the results in the previous example, let's define a few things. We define the training error as

$$\epsilon^2_{\text{train}} = \mathbb{E}[(\hat{Y}_i - Y_{\text{train},i})^2] \tag{151}$$

where the average is taken over different realizations of our data and $\hat{Y}_i$ is our prediction of $\mathbb{E}[Y|X]$. Not the relationship between R^2 and the training error:

$$R^2 \approx 1 - \frac{\epsilon^2_{\text{train}}}{\text{var}(Y_{\text{train},i})}. \tag{152}$$

Similarly, we define the test error as

$$\epsilon^2_{\text{test}} = \mathbb{E}[(\hat{Y}_i - Y_{\text{test},i})^2] \tag{153}$$

It can be shown that

$$\epsilon^2_{\text{test}} = \underbrace{(\mathbb{E}[\hat{Y}_i] - y)^2}_{=\text{bias}} + \underbrace{\text{var}(\hat{Y})}_{=\text{variance}} + \sigma^2_\epsilon \tag{154}$$

The **bias** results from the fact that our model will systematically under or over estimate the y values. For example, if we try to fit an exponentially decaying curve with a straight line, different data sets will give consistent result, but on average they will overestimate the middle of the data and underestimate the ends.

The **variance** variation between our model predictions and the data between different datasets from the same model. For example, if we interpolate every single-point, then a different set of points will cause our curve to change in ways that differ from the data.

Roughly speaking, a biased model will give us consistent but incorrect results, while a low bias high variance model will be correct on average, but our predictions will vary a lot from data set to data set and therefore will not be reliable. **Variance arrises from a model being too complex, while bias arrises from a model not being complex enough.**

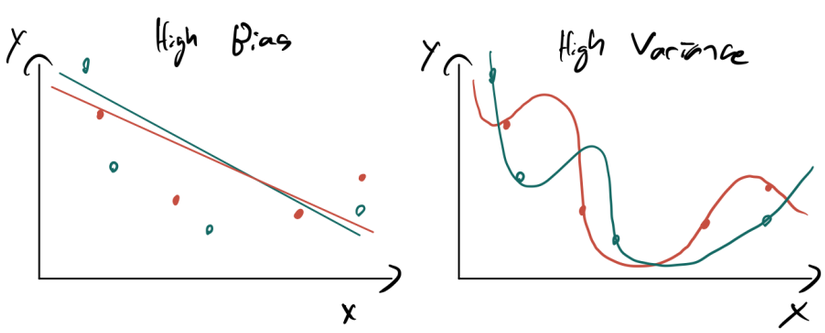

FIGURE 2. Bias and variance illustrated with fits to two different datasets (blue and red) drawn the from the same distribution

Example 32. *Visualizing bias and variance*

Example 33. *Which do we prefer?*

2.3. Information criteria vs. cross validation. It seems like it would be nice to access whether our model is overfitting without splitting the data up, so that (1) we are using all our data and (2) we don't need to refit the model on all the data after cross validation to get our final results. It's also a bit unsatisfying that R^2, which was supposed to measure a models predictive power fails to do that job so miserably as the number of parameters grows. This had motivated a number of alternative metrics for selecting the "best" model which do not relay on out of sample tests. One example is the Akaike information criterion (AIC), which is defined as

$$\text{AIC} = 2k - 2\ln \sum (\hat{Y}_i - Y_i)^2. \tag{155}$$

We prefer models where AIC is small, so the AIC rewards models where $\hat{Y}_i$ and close to Y_i, but penalizes the number of parameters. This seems reasonable, but it's not clear that this is the correct way to penalize parameters. For example, why is AIC linear in k? Why does the k term not depend on how much data we have? There are other information criteria one could use, but they have similar issues. I prefer to simply use cross validation.

The following example illustrates how the AIC doesn't tell us if we are overfitting.

Example 34. *AIC*

3. Additional examples

Example 35. *Earnings data*

CHAPTER 8

Logistic regression

1. Motivation

So far we have considered the linear model:

$$Y = b + \sum a_i X_i + \epsilon \tag{156}$$

where ϵ follows is mean zero Normal random variable. We understand that this can be formulated as the conditional distribution of Y given X:

$$Y|X \sim \text{Normal}\left(b + \sum_i a_i, \sigma_\epsilon\right) \tag{157}$$

We know that we have some formulates for unbiased estimators of b, a_i and σ_ϵ in terms of our data. Specifically, the formulas for the a_i come from the covariances and variances of the predictors and response variables in complicated ways (but are straightforward to compute). After recognizing that we can select both Y and X_i to be function of various variables in our data (by transformations and addition of features), this modeling framework gives the ability to expand our model indefinitely. The remaining limitation is that of the structure of the noise: Since the noise is Gaussian, we are limited to studying only certain times of randomness.

What if we want to predict a binary variable? As an example, let's consider the problem of predicting whether someone supports same sex marriage based on some information about them, such as age and gender. As a "warm up" to get us thinking about this problem, will start with the binary predictor sex (which is restricted to Male or Female is this dataset). Our response variable Y is one if someone supports same sex marriage and zero otherwise. Thus Y is a Bernoulli random variable and thus does not follow a normal distribution regardless of which predictors we condition on. In this case, we can frame the problem of modeling the association between X and Y as two separate inferences of Bernoulli random variables:

$$Y|(X = 0) \sim \text{Bernoulli}(q_0) \tag{158}$$

$$Y|(X = 1) \sim \text{Bernoulli}(q_1) \tag{159}$$

Hence, we can simply break the data up into two groups and estimate q_0 and q_1 as we've done before with Bernoulli random variables.

Alternatively, we can frame this as a regression problem

$$Y|X \sim \text{Bernoulli}(q(X)) \tag{160}$$

where

$$q(X) = q_0(1 - X) + q_1 X = X(q_1 - q_0) + q_0 \tag{161}$$

Structurally, this is similar to the linear regression. Our model for the response variable Y conditioned on X is a distribution in which the parameters depend linearly on X. In the linear regression context, it is the mean of the Normal distribution that depends linearly on X while here it is the chance for $Y = 1$. In fact, if in our data k people support and $N - k$ do not, we compute $\tilde{Y} = k/N$. We could then have a linear regression model (with Normal noise) for $\tilde{Y}$ in terms of X. We will later see that there are some advantages to working directly with Y!

More generally, in order to capture binary noise, we might start with the model

$$Y|X \sim \text{Bernoulli}(q). \tag{162}$$

Here, q is the sole parameter and therefore in order for the distribution of Y to depend on X, q must depend on X. Naively, we might simply take $q(X)$ to be

$$\sum a_i X_i \tag{163}$$

but this has a problem if X are continuous predictors, since we must have $0 < q < 1$. In order to ensure this is the case, we set

$$W = \sum_i a_i X_i \tag{164}$$

and set $q = h(W)$, but we want to select h so that as $W \to \infty$, $q \to 1$, and as $W \to -\infty$, $q \to 0$. The next task is to see how this is done.

2. Logistic model

2.1. The logistic function. We would ideally like to come up with a function $h = h(w)$ that maps w to 0 and 1. The standard choice is

$$q(w) = \text{logit}^{-1}(w) = \frac{1}{1 + e^{-w}} \tag{165}$$

This is called the inverse logistic function between if we solve for w, we get the **logistic function**

$$w = \ln\left(\frac{q}{1-q}\right) \tag{166}$$

To better understand how the slope and intercept b and a effect the plots, let's think about limiting cases. But first, think about the functions e^{-w} and $\text{logit}^{-1}(w)$.

- $e^{-w} = 1$ when $w = 0$. Thus $\text{logit}^{-1}(0) = \frac{1}{1+1} = \frac{1}{2}$.
- If w is a very large positive number, $e^{-w} \approx 0$, so $\text{logit}^{-1}(w) \approx \frac{1}{1+0} = 1$.
- If w is a very large negative number e^{-w} is huge, so $\text{logit}^{-1}(w) \approx \frac{1}{1+\infty} = 0$.

Let's imagine $b = 0$ (there is no intercept).

- If a is very large relative to all values of x, then $w = ax$ will quickly become large for small x and therefore y will very likely be 1 for positive x (and very likely be zero for negative x)
- If a is small relative to all values of x, then $w = ax$ will not change much and the chance that $y = 1$ will be around $1/2$ for most x.

To summarize, our the logistic regression model is

$$Y|X \sim \text{Bernoulli}\left(\frac{1}{1+e^{-b-\sum_i a_i X_i}}\right) \tag{167}$$

Example 36. *Generating data from logistic regression model*

2.2. Fitting logistic regression. We can fit this in statsmodels as shown by the following example, and much of what we've learned turns out to carry over.

Example 37. *Logistic regression in statsmodels*

Example 38. *Logistic regression with real data*

3. Interpretation and visualization

4. Interpreting coefficients

In a logistic regression the meaning of the coefficients is a bit tricky. **This is because their "effect" depends on the value of the predictors.** Let's think about a model with multiple predictors.

Let's think about the intercept first. When the predictor (or all predictors if there are multiple) is zero,

$$P(Y = 1|X = 0) = \frac{1}{1+e^{-b}} \implies b = -\ln\left(\frac{1}{q} - 1\right) \tag{168}$$

We can rearrange terms to get

$$b = \ln\left(\frac{q}{1-q}\right) = \ln\frac{P(Y=1|X=0)}{P(Y=0|X=0)} \tag{169}$$

the expression $(1-q)/q$ is called the odds ratio, so b tells us the log odds ratio to get $y = 1$ when all predictors are zero.

More generally, a_i **tells us how much the log odds ratio changes when we chance** X_i **by** 1 **with all other predictors fixed.** To see this, first note that if the odds are $q = 1/(1+e^{-w})$, the odds ratio is

$$\ln\frac{1/(1+e^{-w})}{1-1/(1+e^{-w})} = \ln\frac{1+e^{-w}}{e^{-w}(1+e^{-w})} = \ln e^{w} = w \tag{170}$$

If we change X_i by 1, then w changes by a_i.

I find this really hard to think about odds ratios. Instead, I think it is easiest to interpret the coefficients when the logistic function is well approximated by a linear function. This happens when

$$w = b + \sum_i a_i X_i = 0. \tag{171}$$

At $w = 0$, $\text{logit}^{-1}(0) = 1/2$ and close to $z = 0$ (between -1 and 1)

$$\text{logit}^{-1}(w) = \frac{1}{1+e^{-w}} \approx \frac{1}{2} + \frac{w}{4} \tag{172}$$

This leads to the **divide by four rule**: When the sum of predictors time coefficients is close to 0, $a_i/4$ represents the difference in the chance that $Y = 1$ between data points for which X_i differs by 1, with all other predictors fixed.

The divide by four rule gives an **upper bound** on how much changing X_i changes the probability for Y to be one. In other words, the actual difference is always less than this.

Example 39. *Homocide victim data*

Exercise 34: *Testing divide by four rule*

5. Model evaluation for logistic regression

It's always useful to have some metric for accessing how much of the variation in the data the model explains the data we used to fit it, even though we know this is not the full story. For linear regression we use R^2. For logistic regression we have something called Pseudo R^2. Like R^2, it tells us, roughly speaking, what fraction of the variation in Y values is explained by the predictors, but in this case we don't have the usual notion of residuals. Instead, we compare how likely it is to see the particular sequence of Y values under the model, vs. how likely it would be to see them if the chance to get $Y = 1$ did not depend on X:

$$\text{pseudo } R^2 = 1 - \frac{\ln(\text{chance to see } Y_1, Y_2, \ldots, Y_n \text{ given our model})}{\ln(\text{chance to see } Y_1, Y_2, \ldots, Y_n \text{ given } x \text{ has no effect})} \tag{173}$$

Why does this make sense? Observe the following facts:

- The chance to see a given sequence of y values of y is between 0 and 1.
- The log of a value between 0 and 1 will be negative, and will be a larger negative number when the chance is smaller.

Thus, if we are much more likely to see the data given our model, the denominator will be closer to 0. If we are less likely to see our data, it will be a large negative number (but always smaller than the denominator).

It tells us how good our model is at predicting the Y values. Notice that if we do bin the data and perform a linear regression, the R^2 we get is generally MUCH larger than the Pseudo R^2. Think about why.

Example 40. *Understanding Pseudo R^2 with simulations*

6. Additional examples

Example 41. *Predicting ground water contamination*

7. More general classification problems

We'd now like to tackle the more general problem predicting a categorical variable. Let's say Y can take on values 0 through $K-1$. To generalize the logistic model, we'd like the probability distribution of Y to have the form

$$P(Y = y|X) = q_i(X). \tag{174}$$

As usual for a categorical variable we only need to define $K-1$ probabilities, with the other being determined by the normalization

$$\sum_{y=0}^{K-1} q_y(X) = 1. \tag{175}$$

Logistic regression corresponds to the case where $K = 2$. In that case, we saw from before that

$$q_1(X) = \frac{1}{1+e^{-W}} \tag{176}$$

and

$$q_0(X) = 1 - q_0(X) = 1 - \frac{1}{1+e^{-W}} = \frac{e^{-W}}{1+e^{-W}} \tag{177}$$

We can equivalently write these as

$$q_1(X) = \frac{1}{Z}, \quad q_0(X) = \frac{e^{-W}}{Z} \tag{178}$$

The values 1 and e^{-W} correspond to **statistical weights** – they are not probabilities, but tell us the relative frequency of these events. Z is a normalization which does not depend on the y values. However, it does depend on X. For a more general regression on a categorical response variable, we will set $q_0 = 1/Z$ and

$$P(Y = y|X) = \frac{e^{-W_y}}{Z}, \quad y = 1, \ldots, K-1. \tag{179}$$

By analogy with logistic regression, we'll have

$$W_y = b_y + \sum_i a_{y,i} X_i. \tag{180}$$

In total, if there are M predictors we have $(M+1)(K-1)$ parameters.

It's important to understand that this so-called multinomial logistic regression model is NOT an intermediate between the logistic regression and the linear regression. If we have many categories that are ordered (say someone's response on a 1 to 10 scale on a survey), it might be better to do a linear regression (understanding that the normal distribution is not a perfect description of the noise). The multinomial logistic regression makes sense when we need to classify things that do not have an obvious ordering.