Week02

August 6, 2021

1 The maximum likelihood approach to finding optimal parameters.

1.1 Introduction

In week 01 many of our diagnostics depended on finding best, or optimal, estimates of the intercept (β_0) and slope (β_1) of our simple linear regression model. But we glossed over *how* those values were computed.

This week we will learn the maximum likelihood (ML) approach to computing optimal parameters. The ML approach is used in numerous statistcial and machine learning models and an approach we will return to throughout this semester.

1.2 Goal

Our goal for week 2 will be to learn: - The approach for finding optimal parameters called the Maximum Likelihood (ML) approach - Mathematical tools to compute optimal parameters (called ML estimators) for a single random variable - Computational aspects of ML - How we can apply ML to simple linear regression to find those optimal intercept and slope parameters

1.3 Probability of Data Set assuming samples are independent and identically distributed (iid)

1.3.1 How parameters determine probabilities and the big idea of Maximum Likelihood

A statistical model typically supposes a way in which data was generated. Most models are associated with a set of **parameters**—values that determine how we assign probabiltiies to a data point. For example, a binomial distribution assigns probabilities to the number of "successes" out of N "trials" for a random variable X using this probability mass function

$$p(X = x|N, \theta) = \binom{N}{x} \theta^x (1-\theta)^{N-x} \tag{1}$$

where θ is the probability of a single success and $\binom{N}{x}$ counts the number of ways x successes can occur among N trials. There are two parameters associated with the Binomial distribution: N and θ, and these parameters determine how the binomial distribution assigns probabilities to values.

For example, when N=10 and $\theta = 0.2$ then the probability of 4 successes is

$$p(X = 4|N = 10, \theta = 0.20) = \binom{10}{4} 0.20^4 (1-0.20)^{10-4} = 0.09 \tag{2}$$

When when N=10 and $\theta = 0.4$ then the probability of 4 successes is

$$p(X = 4|N = 10, \theta = 0.40) = \binom{10}{4} 0.40^4 (1 - 0.40)^{10-4} = 0.25 \tag{3}$$

By changing the parameter θ from 0.20 to 0.40 we changed the probability of observing 4 success from 0.09 to 0.25.

1.3.2 The Big Idea

With the ability to change how we assign probabilities, we can observe a data point and change our parameters to assign higher probabilities to data we observe. Like this: We assume that our single data point is generated by a binomial distribution with parameters N and θ, and lets assume that we know every data point is the number of some success out of 20 trial. Then we can fix $N = 20$ and not need to find this parameter.

But we do need to make a guess at what θ is.

We decide to collect a data point and find that data point is $x_1 = 7$. What should our best guess be?

One approach is to compute the probability our model (binomial distribution) assigns to our data (our single data point) for all potential values of our parameters (for our example, θ). That is, we can plot for every θ the probability that $x = 7$, and plot the pairs

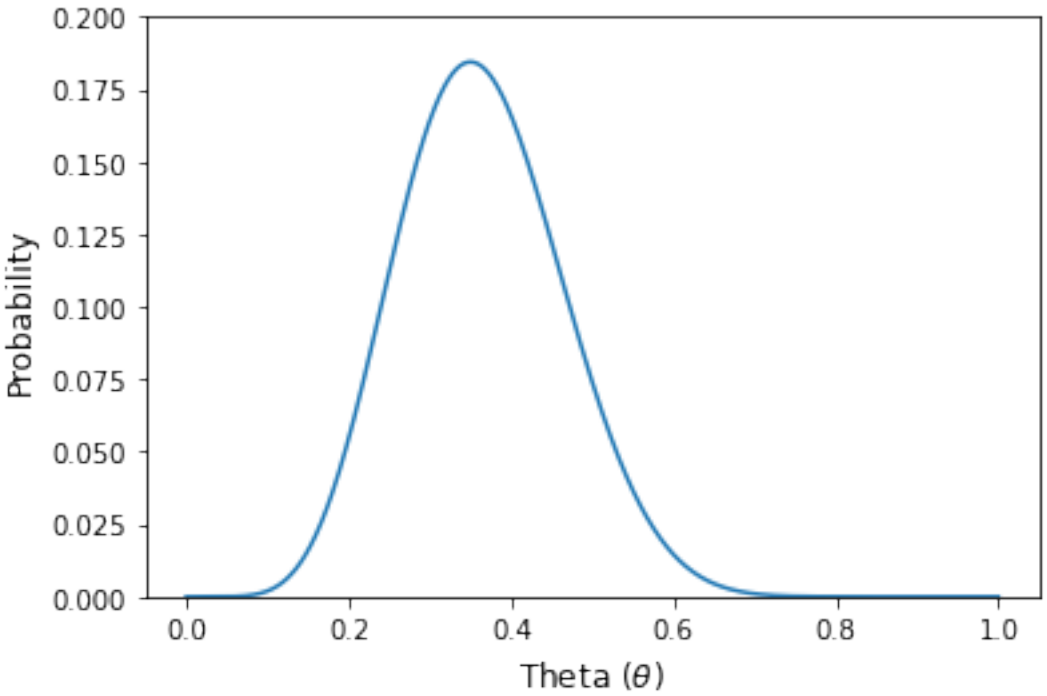
$$(\theta_i, p(X = 7|N = 20, \theta = \theta_i)) \tag{4}$$
$$\tag{5}$$

OR

$$\left(\theta_i, \binom{20}{7} \theta_i^7 (1 - \theta_i)^{10-7}\right) \tag{6}$$

[6]:
```
import scipy
def model(theta):
    return scipy.stats.binom(20,theta).pmf(7)

thetas = np.linspace(0,1,10**3) # Pick 1,000 equally spaced points between the␣
 ↪values 0 and 1

fig,ax = plt.subplots()
ax.plot(thetas, model(thetas))

ax.set_xlabel(r"Theta $(\theta)$",fontsize=12)
ax.set_ylabel("Probability",fontsize=12)

ax.set_ylim(0,0.20)
```

```
plt.show()
```



One way to choose the best possible θ is to pick the θ value above that assigns the highest probability to the data point we observed. This is the idea of maximum likelihood (really).

1.3.3 The probability of more than one data point when they are i.i.d.

Above we saw how parameters change the probability of a single data point. But often we collect much more than a single data point and want to assign a probability

1.4 The Likelihood and LogLikelihood function

1.5 An Example data set:

1.6 (Some) Computational tools for optimizing a loglikelihood function

1.7 Applying ML to Simple Linear Regression

1.8 The Fisher Information Matrix and what it says about our Sampling Distributions