Likelihood and Regression

Author: Nicholas G Reich

This material is part of the ***statsTeachR*** *project*

Today's Lecture

- Likelihood defined
- A simple, coin-flipping example
- Likelihood in the context of regression

These notes are based loosely on Michael Lavine's book Introduction to Statistical Thought, Chapters 2.3-2.4 and 3.2.

Parametric families of distributions

A parametric distribution

- In the analysis of real data, we often are willing to assume that our data come from a distribution whose general form we know, even if we don't know the exact distribution.
- E.g. $X \sim \mathit{Poisson}(\lambda)$ or $Y \sim N(\mu, \sigma^2)$
- Each of the above examples refer to families of distributions, defined or indexed by particular parameter(s).
- In statistics, we try to estimate or learn about the unkown parameter.

The likelihood function

Another look at a pdf

- Probability density functions (pdfs) define the probability of seeing a specific observed value of your random variable, conditional on a parameter.

$$f(X|\theta)$$

- However, we can think about this same function another way, by *conditioning* on the data and looking at the probability taken by different values of the parameter.

$$f(\theta|X) = \ell(\theta)$$

- Remember, the definition of the joint density of observations that we assume to be i.i.d.: if $X_1, X_2, ..., X_n \sim i.i.d. f(x|\theta)$ then

$$f(X_1, \ldots, X_n|\theta) = \prod f(X_i|\theta)$$

Likelihood as evidence

"A wise man ... proportions his belief to his evidence."
-David Hume, Scottish philosopher

We often compare values of the likelihood function as ratios, weighing the evidence for or against particular values of θ.

$$\frac{\ell(\theta_1)}{\ell(\theta_2)} = 1$$

implies we have the same evidence to support either θ_1 or θ_2.

$$\frac{\ell(\theta_1)}{\ell(\theta_2)} > 1$$

implies we have more evidence to support θ_1 over θ_2.

Maximum likelihood estimation

In many settings, there is a unique θ that maximizes $\ell(\theta)$. This value is called the maximum likelihood estimate (a.k.a. the MLE), and is defined

$$\hat{\theta} = argmax_{\theta} \ell(\theta)$$

- MLEs are typically found by taking the derivative of $\log \ell(\theta)$ w.r.t. each parameter and setting equal to zero.
- The likelihood surface is often well behaved, but not always! You could have multiple maxima, a maximum at the boundary of the parameter space, a non-differentiable ℓ, etc...
- MLEs are often intuitive, i.e. for $y_1, y_2, \cdots, y_n \sim N(\mu, \sigma^2)$ the MLE of μ is the sample mean.

Accuracy of estimation

"Doubt is not a pleasant condition, but certainty is an absurd one."
-Voltaire, French writer and philosopher

What other values, in addition to $\hat{\theta}$, have reasonably high likelihood?
We can define a likelihood set (akin to a confidence region) for some value $\alpha \in (0, 1)$, as

$$LS_\alpha := \left\{ \theta : \frac{\ell(\theta)}{\ell(\hat{\theta})} \geq \alpha \right\}$$

- LS are often (but not necessarily) intervals.
- There is no best value of α. Some people like $1/10$. I like $1/8$.
- Typically, as $n \to \infty$ the likelihood becomes more peaked, and the size of LS shrinks.

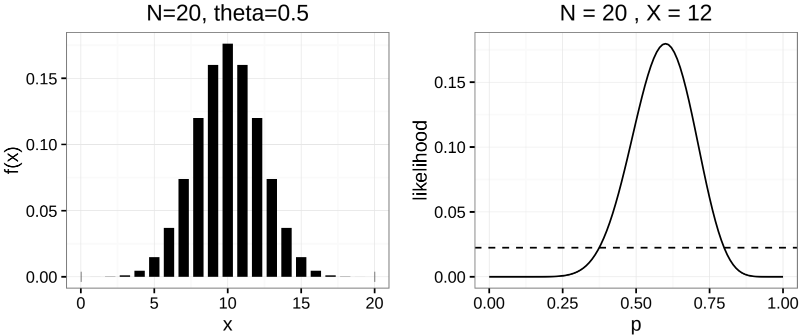
A simple, canonical example: coin-flipping

Let's flip some coins! A plausible statistical model here is for the number of heads (X) when I flip a coin N times

$$X \sim Binomial(N, p)$$

where

$$f(x|p) = \ell(p) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$



A simple, canonical example: coin-flipping

Let's start with three competing hypotheses about my coin and the probability of getting a head:

$$\begin{aligned} H_A : p &= 0.5 \\ H_B : p &= 0 \\ H_C : p &= 1 \end{aligned}$$

```r
source('http://tinyurl.com/coin-likelihood')
coin_lik(x=2, n=4)
```

Numerical optimization of a likelihood function

In R, you can write your own likelihood function and maximize it using one of any number of different functions. For example:

```
ll <- function(p, n, x) -dbinom(x=x, size=n, prob=p, log=TRUE)
## for one-dimensional optimization
optimize(ll, interval=c(0,1), n=10, x=5)

## $minimum
## [1] 0.5
##
## $objective
## [1] 1.402043

## better for multi-dimensional optimization
tmp <- optim(par=list(p=.4), ll, n=10, x=5)

## Warning in optim(par = list(p = 0.4), ll, n = 10, x = 5):
one-dimensional optimization by Nelder-Mead is unreliable:
## use "Brent" or optimize() directly

c(tmp$par, tmp$value)

##        p
## 0.500000 1.402043
```

Likelihood in a regression setting

We have our usual regression model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_p X_{p,i} + \epsilon_i$$

where the ϵ_i are i.i.d. $N(0, \sigma^2)$. So our likelihood function is

$$\ell(\beta_0, \beta_1, \cdots, \beta_p, \sigma) \;=\; \prod_{i=1}^{n} p(y_i | \beta_0, \cdots, \beta_p, \sigma)$$

$$=\; \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} \sum_i \left(y_i - (\beta_0 + \sum \beta_j X_{j,i})\right)^2\right]$$

(Log)-Likelihood in a regression setting

$$
\begin{aligned}
\log \ell(\beta_0, \beta_1, \cdots, \beta_p, \sigma) \quad &= \quad C - n \log \sigma - \\
&\quad\ \frac{1}{2\sigma^2} \sum_i \left(y_i - (\beta_0 + \sum \beta_j X_{j,i}) \right)^2
\end{aligned}
$$

where C is an irrelevant constant. To find the maximum of this likelihood function, we take the derivative of these functions and this gives way to a set of linear equations to solve for the βs. And voila, we have our LSEs again!

Likelihood take-aways

- Likelihood is a flexible and principled framework for evaluating evidence in your data.
- There is strong statistical theory behind likelihood.
- Likelihood is the foundation on which much modern statistical analysis (including most Bayesian analysis) is built.

Finding your own MLEs for regression

Extra credit homework assignment: Take one of the datasets that we have used in class so far and fit a multiple linear regression model (with at least two predictors) using the `optim()` function to obtain maximum likelihood estimators for the regression coefficients and σ. Compare your results to the results from `lm()`.