Week 1: Likelihood Inference

PLSC 40502 - Statistical Models

Welcome!

Course Overview

- Instructor: Anton Strezhnev
- Logistics:
  - Lectures Th - 2:00pm - 4:50pm;
  - 3 Problem Sets (~ 2 weeks)
  - Final paper project (8-12 pages)
  - My office hours: Tuesdays 4pm-6pm (Pick 328)
- What is this course about?
  - *Defining* statistical models via their data-generating process
  - *Estimating* model parameters and conducting *inference*
  - *Interpreting* model output and *evaluating* model quality
- Goals for the course
  - Give you the tools you need to understand descriptive inference via statistical models and comment on other researchers' work.
  - Equip you with an understanding of the fundamentals of likelihood and Bayesian inference to enable you to learn new models that build on these principles.
  - Teach you how to program and implement estimators by yourself!

Course workflow

- Lectures
  - Topics organized by week
  - Lectures are the "course notes" -- readings are the reference manuals.
- Readings
  - Mix of textbooks and papers
  - All readings available digitally on Canvas

Course workflow

- Problem sets (35% of your grade)
    - Meant as a check on your understanding of the material and a way of communicating with me about the course.clear
    - Collaboration is **strongly encouraged** -- you should ask and answer questions on our Ed discussion board.
    - Graded hollistically on a plus/check/minus system.

Course workflow

- Final Project (55% of your grade)
  - The main goal of this class is for you to develop an independent quantitative research project
  - The paper should be in the length and style of a research note (8-12 pages)
  - One well-motivated question + data and analysis (minimize the lit review!)
  - You can collaborate! (1-3 authors per paper).
  - See the syllabus for published examples of the style/method of a paper that fits the aims of this class.
  - Survey data is a great place to ask descriptive questions
  - But feel free to use other sources or ask different types of questions - just talk to me about it!
- Final Project Timeline
  - **February 2nd**: 1 page project memo due
  - **February 29th**: Research presentations in-class (10-15 min. talks + Q&A)
  - **March 7**: Final paper due

Class Requirements

- **Overall**: An interest in learning and willingness to ask questions.
- Assume a background in intro probability and statistics
  - You should be comfortable thinking about basic estimands/estimators + their properties
  - You should be able to interpret a confidence interval for (e.g.) a difference-in-means.
- You should also be familiar with linear regression
  - $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$ should be a familiar expression
  - You should know under what conditions it's unbiased for $E[Y|X]$, and under what conditions it's efficient.
- If you want some review, check out chapters 1-6 of "Regression and Other Stories"

A brief overview

- **Week 1-2:** Introduction to likelihood inference and GLMs
  - Concept of the likelihood, MLE as an estimator + asymptotic properties
  - Binary outcome models, count models, duration models
- **Week 3-4:** Bayesian Inference and Multilevel Models
  - Principles of Bayesian inference -- posteriors, priors, data
  - Quantities of interest: posterior means, credible intervals
  - Estimation via MCMC
  - Application to multilevel regression models
- **Week 5:** Survey data
  - Applying multilevel regression methods to survey data
  - Survey weighting to address non-random sampling.
- **Week 6:** Mixture Models
- **Week 7:** Item response theory and ideal point models
- **Week 8:** Penalized regression and model selection
- **Week 9:** Research presentations + miscellaneous

Defining a statistical model

Regression review

- A very common goal in statistics is to learn about the conditional expectation function $\mathbb{E}[Y|X]$
  - Y_i: Outcome/response/dependent variable
  - X_i: Vector of regressor/independent variables
- "How does the expected value of Y differ across different values of X?"
- Suppose we observe N paired observations of $\{Y_i, X_i\}$.
  - How do we construct a "good" estimator of $\mathbb{E}[Y|X]$?
  - What assumptions do we have to make to get...consistency...unbiasedness...efficiency?
- Consider the ordinary least squares estimator $\hat{\beta}$ which solves the minimization problem:

$$\hat{\beta} = \arg\min_b \sum_{i=1}^{N} (Y_i - X_i b)^2$$

- We can do some algebra and find a closed form solution for this optimization problem

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'Y)$$

Regression review

- **Assumption 1**: Linearity

$$Y = \mathbf{X}\beta + \epsilon$$

- **Assumption 2**: Strict exogeneity of the errors

$$E[\epsilon|\mathbf{X}] = 0$$

- These two imply:
    - Linear CEF

$$\mathbb{E}[Y|\mathbf{X}] = \mathbf{X}\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_k X_k$$

- **Best case**: Our CEF is truly linear (by luck or we have a *saturated* model)
- **Usual case**: We're at least consistent for the *best linear approximation* to the CEF

Regression review

- **Assumption 3**: No perfect collinearity
  - $\mathbf{X}'\mathbf{X}$ is invertible
  - $\mathbf{X}$ has full column rank
- This assumption is needed for *identifiability* -- otherwise no unique solution to the least squares minimization problem exists!
- Fails when one column can be written as a linear combination of the others
  - Or when there are more regressors than observations $k > n$

Regression review

- Under assumptions 1-3, our OLS estimator $\hat{\beta}$ is unbiased and consistent for β
- Let's do a quick proof for unbiasedness

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'Y) \\
&= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'(\mathbf{X}\beta + \epsilon)) \\
&= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\beta + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\epsilon) \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\epsilon)
\end{aligned}$$

- Then we can obtain the conditional expectation of $\mathbb{E}[\hat{\beta}|\mathbf{X}]$

$$\begin{aligned}
\mathbb{E}[\hat{\beta}|\mathbf{X}] &= \mathbb{E}\left[\beta + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\epsilon)\middle|\mathbf{X}\right] \\
&= \mathbb{E}[\beta|\mathbf{X}] + \mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\epsilon)|\mathbf{X}] \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\epsilon|\mathbf{X}] \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'0 \\
&= \beta
\end{aligned}$$

Regression review

- Lastly, by law of total expectation

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[\mathbb{E}[\hat{\beta}|\mathbf{X}]]$$

- Therefore

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[\beta] = \beta$$

- Consistency requires us to show the convergence of $(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\epsilon)$ to 0 in probability as $N \to \infty$.
  - This actually requires *weaker* assumptions: $\mathbb{E}[\mathbf{X}'\epsilon] = 0$ but not necessarily $\mathbb{E}[\epsilon|\mathbf{X}] = 0$.
- But what have we not assumed?
  - Anything about the distribution of the errors!

Regression review

- **Assumption 4** - Spherical errors

$$Var(\epsilon|\mathbf{X}) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}$$

- Benefits
  - Simple, unbiased estimator for the variance of $\hat{\beta}$
  - Completes Gauss-Markov assumptions $\rightsquigarrow$ OLS is BLUE (Best Linear Unbiased Estimator)
- Drawbacks
  - Basically never is true

Regression review

- Good news! We can relax homoskedasticity (but still keep no correlation) and do inference on the variance of $\hat{\beta}$

$$Var(\epsilon|\mathbf{X}) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

- "Robust" standard errors using the "sandwich" estimator - Consistent but not unbiased for the true sampling variance of $\hat{\beta}$
  - Extensions to allow for forms of correlation in the error terms (e.g. "clustering")

Regression review

- **Assumption 5** - Normality of the errors

$$\epsilon|\mathbf{X} \sim \mathcal{N}(0, \sigma^2)$$

- Not necessary even for Gauss-Markov assumptions
- Not needed to do asymptotic inference on $\hat{\beta}$
    - Why? Central Limit Theorem!
- Benefits?
    - Finite-sample inference.

Regression review

- What do we need for OLS to be consistent for the "best linear approximation" to the CEF?
  - Very little!
- What do we need for OLS to be consistent and unbiased for the conditional expectation function?
  - Truly linear CEF
  - But still no assumptions about the outcome distribution!
- What do we need to do inference on $\hat{\beta}$?
  - We almost never assume homoskedasticity because "robust" SE estimators are ubiquitous
  - Even some forms of error correlation are permitted ("cluster" robust SEs)
  - Sample sizes are usually large enough where Central Limit Theorem implies a normal sampling distribution is a reasonable approximation.

Defining a statistical model

- In the regression setting we tried to make as few assumptions about the data-generating process as possible.
  - Our goal is just to estimate and conduct inference on $E[Y|X]$.
- But what if we wanted to make further probabilistic statements about other quantities beyond β?
  - (e.g.) Can we provide a distribution for Y_{n+1}, the "next" observation given X_{n+1}?
  - If we're willing to make more assumptions about the data-generating process, we can do a lot more!
- **Statistical models** specify the data-generating process in terms of *systematic* and *stochastic* components.
  - **Systematic** elements are functions known constants and unknown *parameters*
  - **Stochastic** elements are draws from probability distributions
- We will be primarily working with *parametric* models
  - The data will be assumed to come from a particular family of probability distributions
  - The "structure" of the model is assumed fixed (the number of parameters does not grow with the size of the data).

The linear model

- It is common to see the linear model written in its fully parametric form.
- **Stochastic**:

$$Y_i \sim \text{Normal}(\mu_i, \sigma^2)$$

- **Systematic**:

$$\mu_i = X_i' \beta$$

- What's assumed to be known?
    - $\mathbf{X}$
- What's assumed to have a particular distribution?
    - Y
- We are interested in estimating and conducting inference on the parameters: β and (less importantly) σ^2.

General model notation

- We can specify a broad set of models for Y_i using this framework
- **Stochastic**

$$Y_i \sim f(\theta_i, \alpha)$$

- **Systematic**

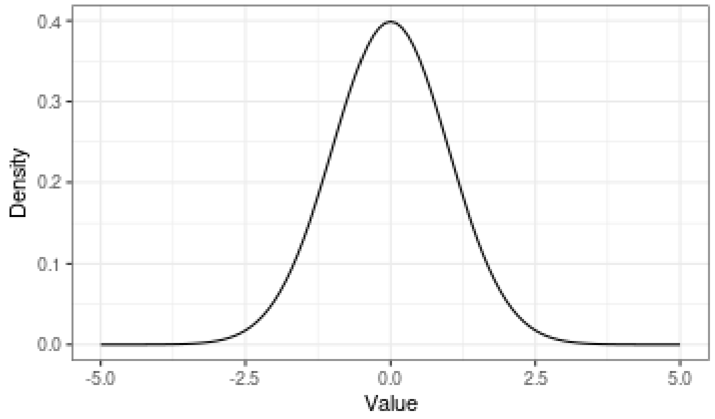
$$\theta_i = g(X_i, \beta)$$

- What are these quantities?
  - Y_i is a random variable
  - $f()$ denotes the distribution of that random variable
  - θ_i and α are parameters of that distribution
  - $g()$ is some function
  - X_i are observed, known constants (e.g. regressors)
  - β are parameters of interest
- We will spend some time with a particular class of models called "Generalized Linear Models" where the systematic component has the form

$$\theta_i = g(X_i'\beta)$$

Types of distributions

- **Normal**

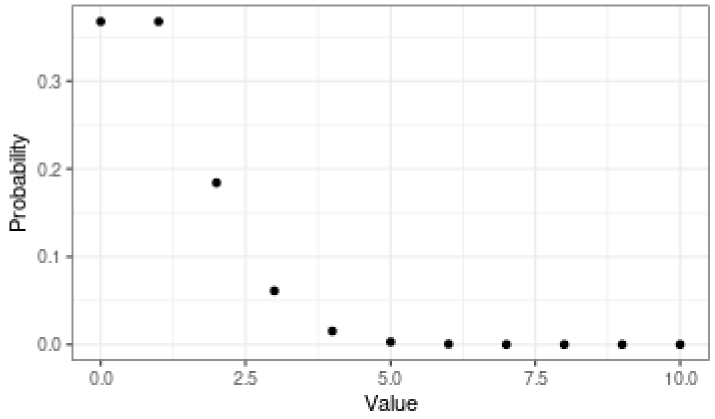


- Continuous on an unbounded support $(-\infty, \infty)$
- Two parameters: Mean μ and Variance σ^2
- Probability Density Function (PDF)

$$f_N(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)\right\}$$

Types of distributions

- **Poisson**



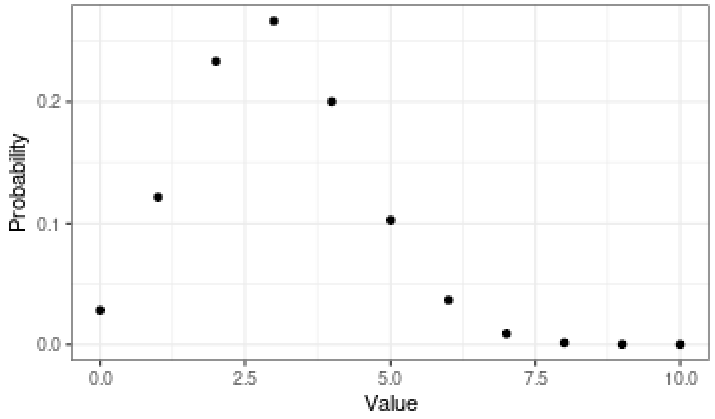
- Discrete, defined on the support of the natural numbers (positive integers + zero)
- Single parameter: Mean and Variance λ

- Probability Density Function (PDF)

$$f(x;\lambda) = \frac{\lambda^x \exp\{-\lambda\}}{x!}$$

Types of distributions

- **Binomial**



- Discrete, defined on the support of integers from $\{0, 1, 2, \ldots, n\}$ (model the sum of repeated i.i.d. coin flips.)
- Two parameters: p probability of success in n trials
  - Special case where $n = 1$ trials is typically called the "Bernoulli"
- Probability Density Function (PDF)

$$f(x; p, n) = \binom{n}{x} p^x (1-p)^{n-x}$$

Likelihood inference

Learning about the unknown

- Suppose that I want to learn about an unobserved parameter from a sample of observations.
  - I *assume* a particular statistical model for the data
  - **Example**: I want to know the level of support for then-President Donald Trump in Wisconsin in 2020 using the 2020 CES
- **Frequentist** approach
  - Unobserved parameters are **fixed constants**
  - Data are **random variables**
- We want to construct an **estimator** that is a *function* of the data and which has desirable properties
  - **Unbiasedness**: The *expected value* of the estimator is equal to the target parameter
  - **Consistency**: As our sample size gets larger, the estimator converges (in probability) to the target parameter.
  - **Asymptotic normality**: In large samples, the distribution of our estimator is normal (ideally with a variance we can estimate as well!)
- Can we come up with a *generic* framework that yields a "good" estimator for a large class of statistical models?

The Likelihood Function

- Consider data $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ which we assume comes from a known distribution with density $f(\mathbf{y}|\theta)$ with unknown parameter θ
  - $f()$ could be Bernoulli, Normal, Poisson, Gamma, Beta, etc... -- in this setting it is a *known* distribution
  - θ is an *unknown* parameter that lies in some space Θ of possible values.
- The **likelihood function** is a function of θ that is evaluated at the observed values of $\mathbf{y}$

$$\mathcal{L}(\theta|\mathbf{y}) = f(\mathbf{y}|\theta)$$

- Given some *input* θ, the likelihood function returns the density of the observed data evaluated at that value.

- The likelihood function is **not** a probability density

  - a pdf is a function that takes x as an input
  - a likelihood is a function that takes θ as an input

The Likelihood Function

- The likelihood function is **not** $f(\theta|\mathbf{y})$
    - That statement doesn't even make sense in a frequentist framework -- parameters are constant
- Even when we (later) move to a *Bayesian* framework $f(\theta|\mathbf{y}) \neq f(\mathbf{y}|\theta)$
- Remember Bayes' Rule:

$$f(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)}{f(\mathbf{y})} \times f(\theta)$$

- Or alternatively, since $f(\mathbf{y})$ is just a normalizing constant, you'll see this written as:

$$\underbrace{f(\theta|\mathbf{y})}_{\text{posterior}} \propto \underbrace{f(\mathbf{y}|\theta)}_{\text{likelihood}} \times \underbrace{f(\theta)}_{\text{prior}}$$

The Likelihood Function

- We often make the assumption that our data are **independently and identically distributed**
  - $y_i \underset{\text{i.i.d}}{\sim} f(y_i|\theta)$
- This allows us to factor the likelihood

$$\mathcal{L}(\theta|\mathbf{y}) = f(\mathbf{y}|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

- Since our eventual goal will be to find an *optimum* of the likelihood, we can apply any monotonic function to it since that preserves maxima/minima.
  - The main one we'll apply is the **logarithm**
  - Why? Because logs turn annoying-to-work-with products into easier-to-work-with sums!
- The log-likelihood is typically denoted $\ell(\theta|\mathbf{y})$

$$\ell(\theta|\mathbf{y}) = \log f(\mathbf{y}|\theta) = \sum_{i=1}^{n} \log f(y_i|\theta)$$

MLE

- We want to come up with an estimator $\hat{\theta}$ for the parameter θ
  - $\hat{\theta}$ is a function of the data (like a sample mean or OLS coefficient)
  - Is there a principled way to pick $\hat{\theta}$ that has provably "good" properties across a wide variety of models?
- The Maximum Likelihood Estimator (MLE) $\hat{\theta}$ is defined as the value of θ that yields the optimum value of the likelihood function

$$\hat{\theta} = \underset{\theta}{\arg\max} \ \log f(\mathbf{y}|\theta)$$

- The MLE has some desirable properties:
  - Under some regularity conditions, the MLE $\hat{\theta}$ is consistent for the true parameter θ
  - It's asymptotically normal: $\sqrt{n}(\hat{\theta}_n - \theta) \to \text{Normal}(0, \frac{1}{\mathcal{I}(\theta)})$
  - It's asymptotically efficient: It's variance approaches the Cramer-Rao lower-bound $\frac{1}{n\mathcal{I}(\theta)}$
  - It's invariant to reparametarization: If $\hat{\theta}$ is an MLE for θ, then $g(\hat{\theta})$ is an MLE for $g(\theta)$

Score

- Before illustrating consistency, we need to define some additional functions of the log-likelihood.
- The first is the **score** or the gradient of $\ell(\theta|\mathbf{y})$ with respect to θ

$$\mathbf{S} = \frac{\partial}{\partial\theta}\log f(\mathbf{y}|\theta)$$

- With i.i.d. data, you'll often see the score written in terms of the score for an individual observation i

$$\mathbf{S}_i = \frac{\partial}{\partial\theta}\log f(y_i|\theta)$$

- At the value of the true parameter θ_0, the expected value of the score is 0 (under regularity conditions)

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(y_i|\theta_0)\right] = \frac{\partial}{\partial\theta}\mathbb{E}[\log f(y_i|\theta_0)] = 0$$

Consistency

- With this definition of the score, we can show consistency of the MLE under i.i.d. observations.
- By the weak law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n}\log f(y_i|\theta) \underset{p}{\rightarrow} \mathbb{E}[\log f(y_i|\theta)]$$

- The MLE is an optimizer of the left-hand side.
    - Therefore it converges in probability to the optimizer of the right-hand side
- And we just showed that the score is 0 at the true value of θ, denoted θ_0
    - Therefore θ_0 is an extremum of the right-hand side (under a few additional regularity conditions)
- Therefore the MLE $\hat{\theta}$ is consistent for the true value θ_0

Consistency

- There are two important conditions for consistency that can be violated in practice
- **Identifiability**
    - No "plateaus" in the log-likelihood.
    - Two different values of θ can't both maximize the log-likelihood

$$f(\mathbf{y}|\theta) \neq f(\mathbf{y}|\theta_0) \; \forall \; \theta \neq \theta_0$$

- **Fixed parameter space**
    - The dimensionality of θ stays fixed and does not depend on n
    - Violated (e.g.) in ideal point models (new legislators mean new ideal points)

Information

- Having established consistency, we're interested in understanding the variance and distribution (asymptotically) of $\hat{\theta}$
- To do this, we need to define a quantity called the **information**: $\mathcal{I}_n$
- This is equivalent to the variance of the score at its optimum

$$\mathcal{I}_n = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{y}|\theta_0)\right)^2\right]$$

- We can show (under some more regularity conditions) that this is equivalent to the negative expectation of the second-order partial derivative (Hessian) of the log-likelihood

$$\mathcal{I}_n = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{y}|\theta_0)\right]$$

- **Intuitively**: Captures the curvature of the log-likelihood at its maximum

Information

- You'll sometimes see the information written in terms of the information from a single observation

$$\mathcal{I} = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(y_i|\theta_0)\right]$$

- Under our i.i.d. assumption, we can write

$$\mathcal{I}_n = n\mathcal{I}$$

- This matters a bit for how we write/show consistency of the MLE

Information

- Under our i.i.d. assumption (+ regularity conditions), we can show not only consistency but convergence in distribution to a normal

$$\sqrt{n}(\hat{\theta} - \theta) \underset{d}{\rightarrow} \text{Normal}(0, \mathcal{I}^{-1})$$

- So in large samples, a good approximation of the variance of $\hat{\theta}$ is $\mathcal{I}_n^{-1}$
  - An additional result: The inverse of the information is the lowest possible variance of any unbiased estimator (**Cramér-Rao lower bound**)
  - MLE achieves this bound asymptotically
- How do we do inference then? Same way as before -- plug in a consistent estimator of $\mathcal{I}_n^{-1}$.
  - Calculate the hessian of the log-likelihood at the MLE and take the inverse of its negative.

Example

Example: Bernoulli

- Suppose I was interested in learning about the proportion of residents in Wisconsin who approved of then-President Donald Trump's performance in 2020.
    - I observe a sample of n respondents and observe an approve $(y_i = 1)$ or disapprove $(y_i = 0)$ response.
    - Ignore, for this example, the sample weights and assume we have a true simple random sample from the target populations.
- The data generating process:
    - *Stochastic*: $y_i \sim \text{Bernoulli}(\pi)$
    - *Systematic*: $\pi \in (0, 1)$
- Let's derive the (log)likelihood!

Example: Bernoulli

- Under our assumption of i.i.d. observations:

$$\mathcal{L}(\pi|\mathbf{y}) = f(\mathbf{y}|\pi) = \prod_{i=1}^{n} f(y_i|\pi)$$

- The log-likelihood is

$$\ell(\pi|\mathbf{y}) = \sum_{i=1}^{n} \log f(y_i|\pi)$$

- Since y_i is Bernoulli, the PMF is:

$$f(y_i|\pi) = \pi^{y_i}(1-\pi)^{1-y_i}$$

- Plugging back into the log-likelihood

$$\ell(\pi|\mathbf{y}) = \sum_{i=1}^{n} \log\left(\pi^{y_i}(1-\pi)^{1-y_i}\right)$$

Example: Bernoulli

- Log of the product is the sum of the logs

$$\ell(\pi|\mathbf{y}) = \sum_{i=1}^{n} \log\left(\pi^{y_i}\right) + \log\left((1-\pi)^{1-y_i}\right)$$

- Properties of logs: $\log(a^b) = b\log(a)$

$$\ell(\pi|\mathbf{y}) = \sum_{i=1}^{n} y_i \log(\pi) + (1-y_i)\log(1-\pi)$$

- We'll simplify this further later, but let's take a look at the shape of this likelihood function w.r.t. its input π

Example: Bernoulli

- Read in the data

```r
approval <- read_csv("data/cces2020_trump_approval_WI_TX.csv")
approval <- approval %>% filter(!is.na(trumpapprove)) # Drop missing
approval_WI <- approval %>% filter(inputstate == 55) # Subset down to Wisconsin
approval_TX <- approval %>% filter(inputstate == 48) # Also get TX for comparison
```

- Write the likelihood function in code

```r
bern_lik <- function(pi, y){
  return(sum(y*log(pi) + (1-y)*log(1-pi)))
}
```

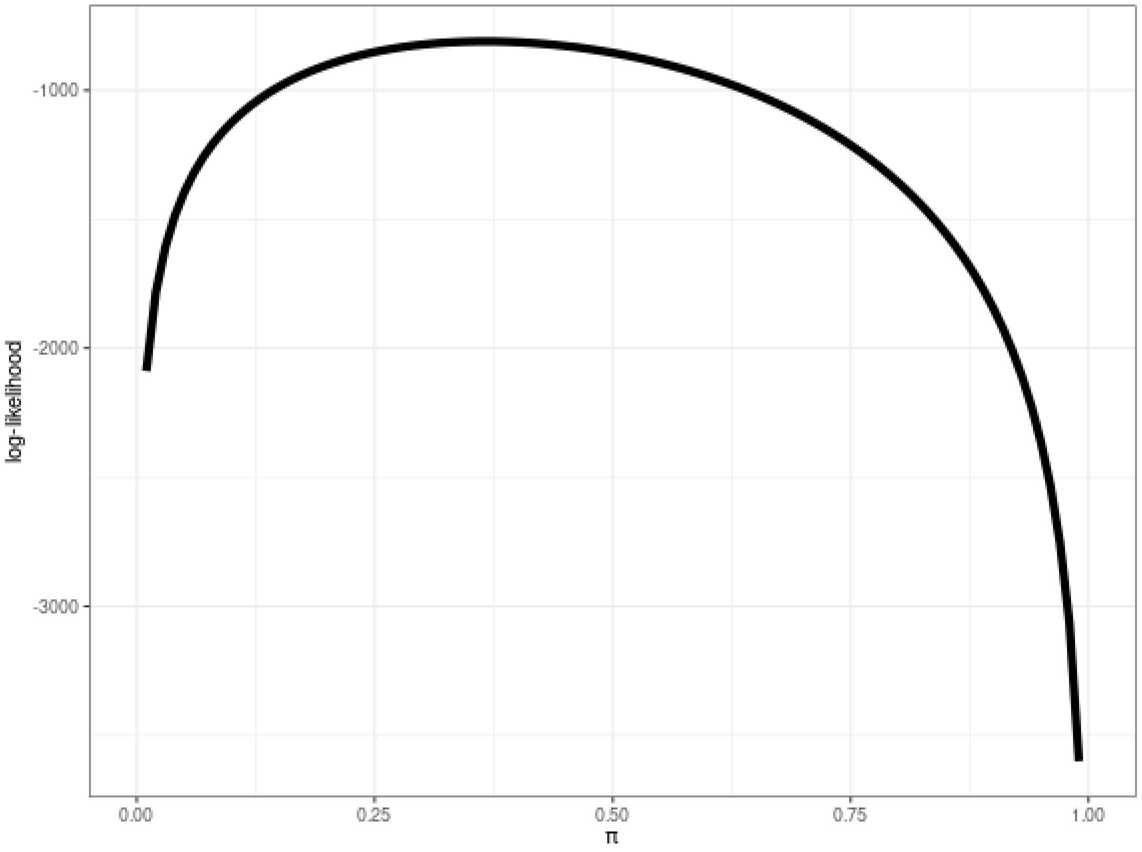
- Quick summary of the data

```r
table(approval_WI$trumpapprove)
```

```
##
##   0   1
## 781 452
```

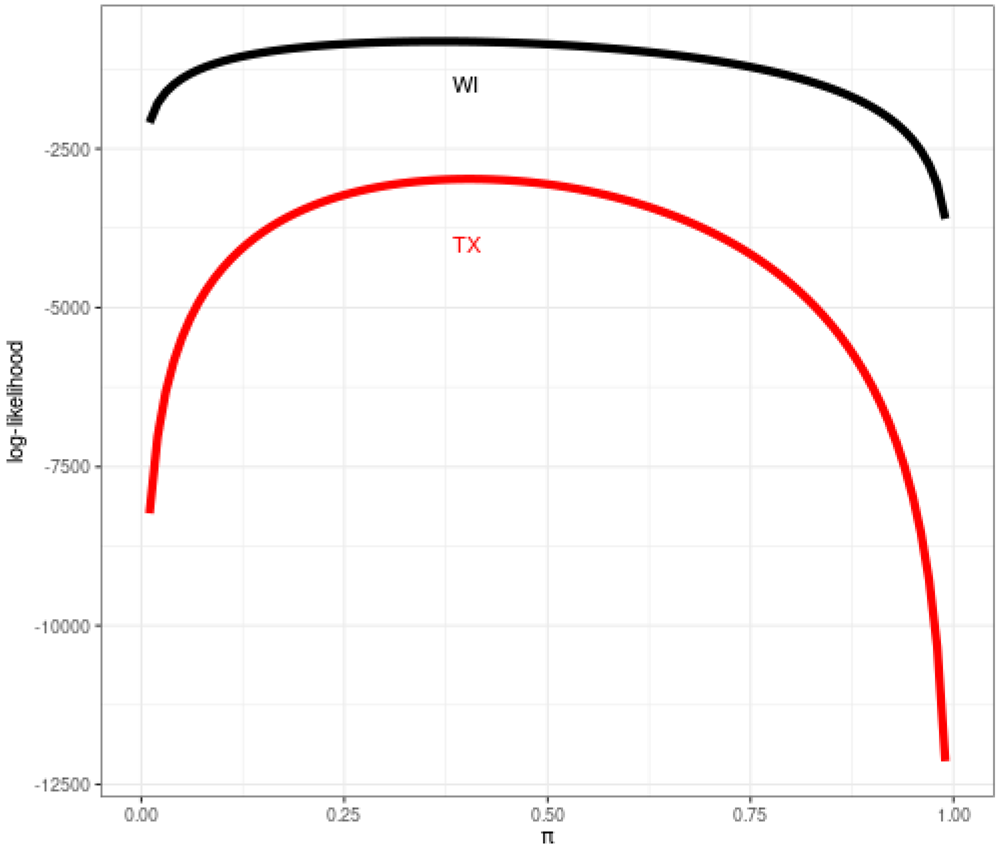
Visualizing the likelihood

- For Wisconsin



Visualizing the likelihood

- For Wisconsin and Texas



Numerical optimization

- Often the MLEs can be obtained via closed-form solutions, but in most applications, the answer has to be obtained numerically.
- R has a built-in numerical optimizer: `optim` that implements some standard algorithms
  - But *tons* of other packages: https://cran.r-project.org/web/views/Optimization.html
- Let's use `optim()` to calculate the MLE

```
# Pass our likelihood through to the optimizer
mle_wi_optim <- optim(.5, fn=bern_lik,
                      y=approval_WI$trumpapprove,
                      method = "BFGS",
                      control=list(fnscale=-1),
                      hessian=T)
```

Numerical optimization

- What's the output of `optim()`?

```
mle_wi_optim
```

```
## $par
## [1] 0.367
##
## $value
## [1] -810
##
## $counts
## function gradient
##       24        6
##
## $convergence
## [1] 0
##
## $message
## NULL
##
## $hessian
##       [,1]
## [1,] -5310
```

Numerical optimization

- Our MLE point estimate is:

```
point_mle <- mle_wi_optim$par
point_mle
```

```
## [1] 0.367
```

- A consistent estimator of the asymptotic variance is the inverse of the negative hessian

```
var_mle <- solve(-mle_wi_optim$hessian)
sqrt(var_mle) # Standard error
```

```
##         [,1]
## [1,] 0.0137
```

- So our asymptotic 95% CI is

```
c(point_mle - abs(qnorm(.025))*sqrt(var_mle),
  point_mle + abs(qnorm(.025))*sqrt(var_mle))
```

```
## [1] 0.340 0.393
```

Analytical optimization

- The Bernoulli case is pretty simple, so let's try to find an analytical expression for the MLE. We left off at

$$\ell(\pi|\mathbf{y}) = \sum_{i=1}^{n} y_i \log(\pi) + (1 - y_i) \log(1 - \pi)$$

- To find the optimum, we first take the partial derivative with respect to π and set it equal to zero

$$\frac{\partial}{\partial \pi}\ell(\pi|\mathbf{y}) = \frac{\partial}{\partial \pi}\sum_{i=1}^{n} y_i \log(\pi) + (1 - y_i) \log(1 - \pi)$$

- Derivatives and sums

$$\frac{\partial}{\partial \pi}\ell(\pi|\mathbf{y}) = \sum_{i=1}^{n} \left[\frac{\partial}{\partial \pi} y_i \log(\pi)\right] + \left[\frac{\partial}{\partial \pi}(1 - y_i) \log(1 - \pi)\right]$$

Analytical optimization

- Chain rule and product rules

$$\frac{\partial}{\partial \pi}\ell(\pi|\mathbf{y}) = \sum_{i=1}^{n} \frac{y_i}{\pi} - \frac{1-y_i}{1-\pi}$$

- Split into 3 terms and pull out constants

$$\frac{\partial}{\partial \pi}\ell(\pi|\mathbf{y}) = \frac{1}{\pi}\sum_{i=1}^{n} y_i + \frac{1}{1-\pi}\sum_{i=1}^{n} y_i - \frac{n}{1-\pi}$$

- Set equal to 0 and solve for π

$$0 = \frac{1}{\hat{\pi}}\sum_{i=1}^{n} y_i + \frac{1}{1-\hat{\pi}}\sum_{i=1}^{n} y_i - \frac{n}{1-\hat{\pi}}$$

Analytical optimization

- Rearrange terms

$$\frac{n}{1-\hat{\pi}} = \frac{1}{\hat{\pi}} \sum_{i=1}^{n} y_i + \frac{1}{1-\hat{\pi}} \sum_{i=1}^{n} y_i$$

- Multiply through by $1 - \hat{\pi}$

$$n = \frac{1-\hat{\pi}}{\hat{\pi}} \sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i$$

- Cancel terms

$$n = \frac{1}{\hat{\pi}} \sum_{i=1}^{n} y_i$$

Analytical optimization

- Divide by n, multiply by $\hat{\pi}$

$$\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Check that the second derivative is negative
  - Here it's strictly negative over the domain of π

$$\frac{\partial^2}{\partial \pi^2}\ell(\pi|\mathbf{y}) = \sum_{i=1}^{n} -\frac{y_i}{\pi^2} - \frac{1-y_i}{(1-\pi)^2}$$

$$\frac{\partial^2}{\partial \pi^2}\ell(\pi|\mathbf{y}) = -\frac{1}{\pi^2}\sum_{i=1}^{n} y_i - \frac{1}{(1-\pi)^2}\sum_{i=1}^{n}(1-y_i)$$

- All that work and it turns out the MLE is just the **sample mean**!
  - Won't always be the case, but here the MLE is also *unbiased* because of what we know about the sample mean.

CRLB

- Lastly, can we can also show that this particular estimator reaches the Cramer-Rao Lower Bound?
- Start with the known variance of the sample mean $\hat{\pi}$

$$Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n}$$

- Next, let's write the Fisher information

$$\mathcal{I}_n = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{y}|\theta_0)\right]$$

- From above

$$\mathcal{I}_n = -\mathbb{E}\left[-\frac{1}{\pi^2}\sum_{i=1}^{n} y_i - \frac{1}{(1-\pi)^2}\sum_{i=1}^{n}(1-y_i)\right]$$

$$\mathcal{I}_n = \mathbb{E}\left[\frac{1}{\pi^2}\sum_{i=1}^{n} y_i + \frac{1}{(1-\pi)^2}\sum_{i=1}^{n}(1-y_i)\right]$$

CRLB

- Pulling out the constants and applying linearity

$$\mathcal{I}_n = \frac{1}{\pi^2}\sum_{i=1}^{n}\mathbb{E}[y_i] + \frac{1}{(1-\pi)^2}\sum_{i=1}^{n}\mathbb{E}[(1-y_i)]$$

- Under the model

$$\mathcal{I}_n = \frac{n\pi}{\pi^2} + \frac{n(1-\pi)}{(1-\pi)^2}$$

- Cancelling and factoring

$$\mathcal{I}_n = n\left(\frac{1}{\pi} + \frac{1}{(1-\pi)}\right)$$

CRLB

- Adding the fractions

$$\mathcal{I}_n = \frac{n}{\pi(1-\pi)}$$

The CRLB (for an unbiased estimator) is the inverse of the Fisher information

$$\frac{1}{\mathcal{I}_n} = \frac{\pi(1-\pi)}{n}$$

which is equivalent to the variance of our estimator. This estimator attains the CRLB (it's a minimum-variance unbiased estimator)!

Conclusion

- **Statistical models**
  - Describe the data-generating process in terms of *systematic* and *stochastic* components
  - In a parametric model, we assume the data $\mathbf{y}$ come from a known *distribution* with unknown parameters θ
- **Likelihood**
  - A function of θ evaluated at the observed data $\mathbf{y}$ equal to $f(\mathbf{y}|\theta)$
  - How "likely" are my observed results in a world where the true DGP parameter were θ.
  - Not $f(\theta|\mathbf{y})$. That quantity only makes sense in the Bayesian context and requires us to formulate a prior $f(\theta)$
- **Maximum Likelihood Estimator**
  - A *technique* to come up with a **good** estimator in any case where we can write down a (well-behaved) likelihood function for a DGP
  - Still a "function" of the data, but that function has useful properties:
  - Consistency, Asymptotic Normality, Asymptotic Efficiency