Advanced Regression

Often we are interested in making inferences and predictions from data, either by (1) estimating particular meaningful parameters of models or (2) finding best fitting model that we can then manipulate to produce useful outputs such as predictions or counterfactual estimates. Focus on what is done when linear models are not appropriate and may produce misleading estimates. Generalized linear model and maximum likelihood methods as essential tools all statistics students should understand. Examination to shift gears to explore predictive modeling techniques that have been ubiquitous in machine learning ~~literature~~ [practice] in recent years~~, with special attention to regularization and kernelized methods.~~

Multinomial Regression

Robust Regression

Dirichlet Regression

PolyMARS Regression

Isotonic Regression

Multiple Linear Regression

SVM Regression

Polynomial Regression

ElasticNet Regression

Quantile Regression

PCA Regression

LASSO Regression

Logistic Regression

Ridge Regression

PLS Regression

Cox Regression

Negative Binomial Regression

Quasi-Poisson Regression

Ordinal Regression

Poisson Regression

Beta Regression

Advanced Regression

simply put, regression is used to model the relationship between a dependent variable and one or more independent variables

previously you have spent (lots) of time focusing on multiple linear regression and variants concerned with inferences and predictions on continuous response variables

Classification

assigning observations to a category or class through predicting qualitative (categorical) response

often the methods used for classification first predict the probability of each of the categories of the qualitative variable as the basis for making the classification

we will begin by considering a binary response

Logistic Regression

standard (often first) way to model binary outcomes

most common way to make inferences on coefficients for binary outcome to understand something about the world

Design-based Inference

- randomness comes from a controlled study design
- examples: randomized experiments, survey sampling
- reliable framework for causal inference
- not always possible

Model-based Inference

- how do we set parameters of a pre-specified model to best accord with the data?
- randomness comes from an assumed probability model
- good for predictive inference, characterizing data; can be used for causal inference only under difficult assumptions

*here we focus on model-based inference, but many projects involve aspects of both

OLS targeted the Conditional Expectation Function,

$$\mathbb{E}[Y_i|X_i] = X_i^\top \beta$$

OLS targeted the Conditional Expectation Function,

$$\mathbb{E}[Y_i|X_i] = X_i^\top \beta$$

Note focus on expectation only, not on $p(Y_i|X_i)$

OLS targeted the Conditional Expectation Function,

$$\mathbb{E}[Y_i|X_i] = X_i^\top \beta$$

Note focus on expectation only, not on $p(Y_i|X_i)$

Alternatively:

- Assume Y_i drawn from a distribution that depends on X_i and parameters:

$$Y_i|X_i \sim p(Y_i|\theta_i = g(X_i, \theta))$$

- then choose the θ that makes the observed sample most likely

OLS targeted the Conditional Expectation Function,

$$\mathbb{E}[Y_i|X_i] = X_i^\top \beta$$

Note focus on expectation only, not on $p(Y_i|X_i)$

Alternatively:

- Assume Y_i drawn from a distribution that depends on X_i and parameters:

$$Y_i|X_i \sim p(Y_i|\theta_i = g(X_i, \theta))$$

- then choose the θ that makes the observed sample most likely

Downsides: you'll have to assume

- family of the distribution
- functional form relating X_i to θ_i

OLS targeted the Conditional Expectation Function,

$$\mathbb{E}[Y_i|X_i] = X_i^\top \beta$$

Note focus on expectation only, not on $p(Y_i|X_i)$

Alternatively:

- Assume Y_i drawn from a distribution that depends on X_i and parameters:

$$Y_i|X_i \sim p(Y_i|\theta_i = g(X_i, \theta))$$

- then choose the θ that makes the observed sample most likely

Downsides: you'll have to assume

- family of the distribution
- functional form relating X_i to θ_i

Upside:

- unified approach to model wide variety of outcome variables
- you get a best unbiased estimator (when the assumptions are right)

Really the whole idea of maximum likelihood comes down to:

- Assume the outcome (Y_i) generated by distribution that depends on the X_i and some parameters.

Really the whole idea of maximum likelihood comes down to:

- Assume the outcome (Y_i) generated by distribution that depends on the X_i and some parameters.
  - CoinLandsHeads $\sim$ *Bern*(π)
  - $Y_i \sim N(X_i^\top \beta, \sigma^2)$
  - *TrumpWinsInStateX* $\sim$ *Bern*($\pi_i = g(X_i, \beta)$)

Really the whole idea of maximum likelihood comes down to:

- Assume the outcome (Y_i) generated by distribution that depends on the X_i and some parameters.
  - CoinLandsHeads $\sim$ *Bern*(π)
  - $Y_i \sim N(X_i^\top \beta, \sigma^2)$
  - *TrumpWinsInStateX* $\sim$ *Bern*($\pi_i = g(X_i, \beta)$)
- We hope to learn something about the world from knowing the parameters of such a model.
- Guess values of the parameters and see how likely the observed data would be had those been the true parameters
- Choose the value of the parameters that maximize the likelihood of the observed data.

- Random variable Z has pdf $p(Z)$ or pmf $p(Z)$. (Rather than $f_Z(Z)$)
- Outcome Y with $dim(Y) = 1$, covariates X with $dim(X) = P$.
- A underlying joint distribution, $p(X, Y)$
- Samples $\{X_i, Y_i\}$ drawn independently from $p(X, Y)$, $i = 1, ..., N$.
- We use the sampled data, $\mathcal{D} = \{X, Y\}_i^N$ to learn about $p(X, Y)$
- Note that since X_1, X_3 or any X_i is drawn from $p(X, Y)$, we can speak of $\mathbb{E}[X]$ or $\mathbb{E}[X_i]$ interchangeably.

Why not linear probability model (LPM), $\mathbb{E}[Y_i|X_i] = X_i^\top \beta$?

Why not linear probability model (LPM), $\mathbb{E}[Y_i|X_i] = X_i^\top \beta$?

- $X_i^\top \beta$ could be outside $[0, 1]$
- Errors depend on $X_i \Rightarrow$ always heteroskedastic

Why not linear probability model (LPM), $\mathbb{E}[Y_i|X_i] = X_i^\top \beta$?

- $X_i^\top \beta$ could be outside $[0, 1]$
- Errors depend on $X_i \Rightarrow$ always heteroskedastic

But we *really* think $Y_i \sim Bern(\pi_i)$, where

- π_i (could) depend on X_i somehow: $Y_i|X_i \sim Bern(\pi_i)$

Why not linear probability model (LPM), $\mathbb{E}[Y_i|X_i] = X_i^\top \beta$?

- $X_i^\top \beta$ could be outside $[0, 1]$
- Errors depend on $X_i \Rightarrow$ always heteroskedastic

But we *really* think $Y_i \sim Bern(\pi_i)$, where

- π_i (could) depend on X_i somehow: $Y_i|X_i \sim Bern(\pi_i)$
- Alternative notations you will see:
  - $Y_i|X_i \sim p(Y_i|X_i, \theta)$, or
  - $Y_i|X_i \sim p(Y_i|\theta_i)$

Suppose we had $Y_i \sim Bern(\pi)$, with fixed π.

Suppose we had $Y_i \sim Bern(\pi)$, with fixed π.

Let's draw some Y_i independently.

Suppose we had $Y_i \sim Bern(\pi)$, with fixed π.

Let's draw some Y_i independently.

After one observation,

- $Pr(Y_i = 1)$?

Suppose we had $Y_i \sim Bern(\pi)$, with fixed π.

Let's draw some Y_i independently.

After one observation,

- $Pr(Y_i = 1)$? π

Suppose we had $Y_i \sim Bern(\pi)$, with fixed π.

Let's draw some Y_i independently.

After one observation,

- $Pr(Y_i = 1)$? π
- $Pr(Y_i = 0)$?

Suppose we had $Y_i \sim Bern(\pi)$, with fixed π.

Let's draw some Y_i independently.

After one observation,

- $Pr(Y_i = 1)$? π
- $Pr(Y_i = 0)$? $1 - \pi$

Suppose we had $Y_i \sim Bern(\pi)$, with fixed π.

Let's draw some Y_i independently.

After one observation,

- $Pr(Y_i = 1)$? π
- $Pr(Y_i = 0)$? $1 - \pi$
- $Pr(Y_i = y)$?

Suppose we had $Y_i \sim Bern(\pi)$, with fixed π.

Let's draw some Y_i independently.

After one observation,

- $Pr(Y_i = 1)$? π
- $Pr(Y_i = 0)$? $1 - \pi$
- $Pr(Y_i = y)$? $\pi^y(1 - \pi)^{(1-y)}$

Suppose we had $Y_i \sim Bern(\pi)$, with fixed π.

Let's draw some Y_i independently.

After one observation,

- $Pr(Y_i = 1)$? π
- $Pr(Y_i = 0)$? $1 - \pi$
- $Pr(Y_i = y)$? $\pi^y(1 - \pi)^{(1-y)}$

After two observations,

Suppose we had $Y_i \sim Bern(\pi)$, with fixed π.

Let's draw some Y_i independently.

After one observation,

- $Pr(Y_i = 1)$? π
- $Pr(Y_i = 0)$? $1 - \pi$
- $Pr(Y_i = y)$? $\pi^y(1-\pi)^{(1-y)}$

After two observations,

$$\begin{aligned} Pr(Y_1 = y_1, Y_2 = y_2) &= Pr(Y_1 = y_1)Pr(Y_2 = y_2) \\ &= \pi^{y_1}(1-\pi)^{(1-y_1)}\pi^{y_2}(1-\pi)^{(1-y_2)} \\ &= \pi^{\sum_i y_i}(1-\pi)^{\sum_i (1-y_1)} \end{aligned}$$

...after N observations:

$$Pr(Y_1 = y_1, ..., Y_N = y_n) = \prod_i Pr(Y_i = y_i)$$
$$= \prod_i \pi^{y_i}(1 - \pi)^{(1-y_i)}$$

...after N observations:

$$Pr(Y_1 = y_1, ..., Y_N = y_n) = \prod_i Pr(Y_i = y_i)$$
$$= \prod_i \pi^{y_i}(1 - \pi)^{(1-y_i)}$$

We can call this quantity:

...after N observations:

$$Pr(Y_1 = y_1, ..., Y_N = y_n) = \prod_i Pr(Y_i = y_i)$$
$$= \prod_i \pi^{y_i}(1 - \pi)^{(1-y_i)}$$

We can call this quantity:

- the "probability of the data" given π
- $p(\boldsymbol{y}|\pi)$, where $\boldsymbol{y} = [Y_1, ..., Y_N]$
- The likelihood, $L(\pi|\boldsymbol{y})$

...after N observations:

$$Pr(Y_1 = y_1, ..., Y_N = y_n) = \prod_i Pr(Y_i = y_i)$$
$$= \prod_i \pi^{y_i}(1-\pi)^{(1-y_i)}$$

We can call this quantity:

- the "probability of the data" given π
- $p(\boldsymbol{y}|\pi)$, where $\boldsymbol{y} = [Y_1, ..., Y_N]$
- The likelihood, $L(\pi|\boldsymbol{y})$

$L(\pi|\boldsymbol{y})$ is the same as $p(\boldsymbol{y}|\pi)$, but arguments reversed

...after N observations:

$$Pr(Y_1 = y_1, ..., Y_N = y_n) = \prod_i Pr(Y_i = y_i)$$
$$= \prod_i \pi^{y_i}(1 - \pi)^{(1-y_i)}$$

We can call this quantity:

- the "probability of the data" given π
- $p(\boldsymbol{y}|\pi)$, where $\boldsymbol{y} = [Y_1, ..., Y_N]$
- The likelihood, $L(\pi|\boldsymbol{y})$

$L(\pi|\boldsymbol{y})$ is the same as $p(\boldsymbol{y}|\pi)$, but arguments reversed

Why? Data are fixed, we're going to vary π

What choice of π maximizes the chances of observing $\boldsymbol{y}$?

$$\hat{\pi}_{MLE} = \underset{\pi \in [0,1]}{argmax} \prod_i L(\pi | Y_i)$$

$$= \underset{\pi \in [0,1]}{argmax} \prod_i \pi^{Y_i} (1 - \pi)^{(1 - Y_i)}$$

What choice of π maximizes the chances of observing $\boldsymbol{y}$?

$$\begin{aligned}\hat{\pi}_{MLE} &= \underset{\pi \in [0,1]}{argmax} \prod_i L(\pi | Y_i) \\ &= \underset{\pi \in [0,1]}{argmax} \prod_i \pi^{Y_i}(1-\pi)^{(1-Y_i)}\end{aligned}$$

Easier to instead maximize $log(L(\pi|y))$,

$$log(L(\pi|\boldsymbol{y})) = \ell(\pi|\boldsymbol{y}) = \sum_i Y_i log(\pi) + (1 - Y_i) log(1-\pi)$$

The “maximum likelihood” estimate of π, $\hat{\pi}_{MLE}$ is thus:

$$\hat{\pi}_{MLE} = \underset{\hat{\pi} \in [0,1]}{argmax} \sum_i Y_i log(\hat{\pi}) + (1 - Y_i) log(1 - \hat{\pi})$$

The “maximum likelihood” estimate of π, $\hat{\pi}_{MLE}$ is thus:

$$\hat{\pi}_{MLE} = \underset{\hat{\pi} \in [0,1]}{argmax} \sum_i Y_i log(\hat{\pi}) + (1 - Y_i) log(1 - \hat{\pi})$$

The first-derivative of $\ell(\pi)$ is called the score, $S(\pi)$.

The "maximum likelihood" estimate of π, $\hat{\pi}_{MLE}$ is thus:

$$\hat{\pi}_{MLE} = \underset{\hat{\pi} \in [0,1]}{argmax} \sum_i Y_i log(\hat{\pi}) + (1 - Y_i) log(1 - \hat{\pi})$$

The first-derivative of $\ell(\pi)$ is called the score, $S(\pi)$.

Maximum likelihood occurs where $S(\pi) = 0$

$$S(\pi) = \frac{d\ell(\pi|\boldsymbol{y})}{d\pi} = 0$$

$$0 = \sum_i \frac{Y_i}{\pi} - \sum_i \frac{(1 - Y_i)}{1 - \pi}$$

$$\frac{1}{\pi} \sum_i Y_i = \frac{1}{1 - \pi} (\sum_i Y_i - N)$$

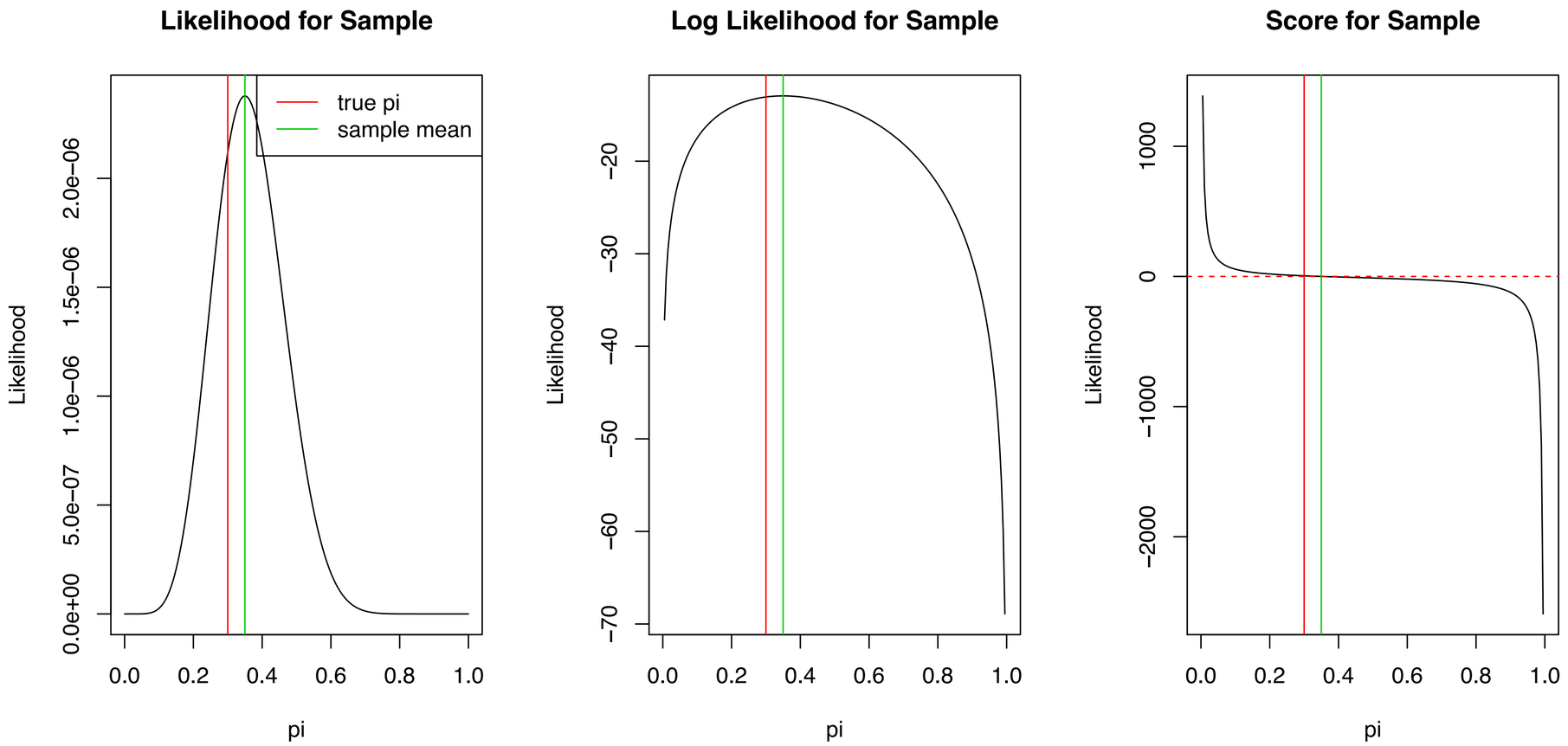
$$\frac{1 - \pi}{\pi} + 1 = \frac{N}{\sum_i Y_i}$$

$$\frac{1}{\pi} = \frac{N}{\sum_i Y_i}$$

$$\pi = \frac{\sum_i Y_i}{N}$$

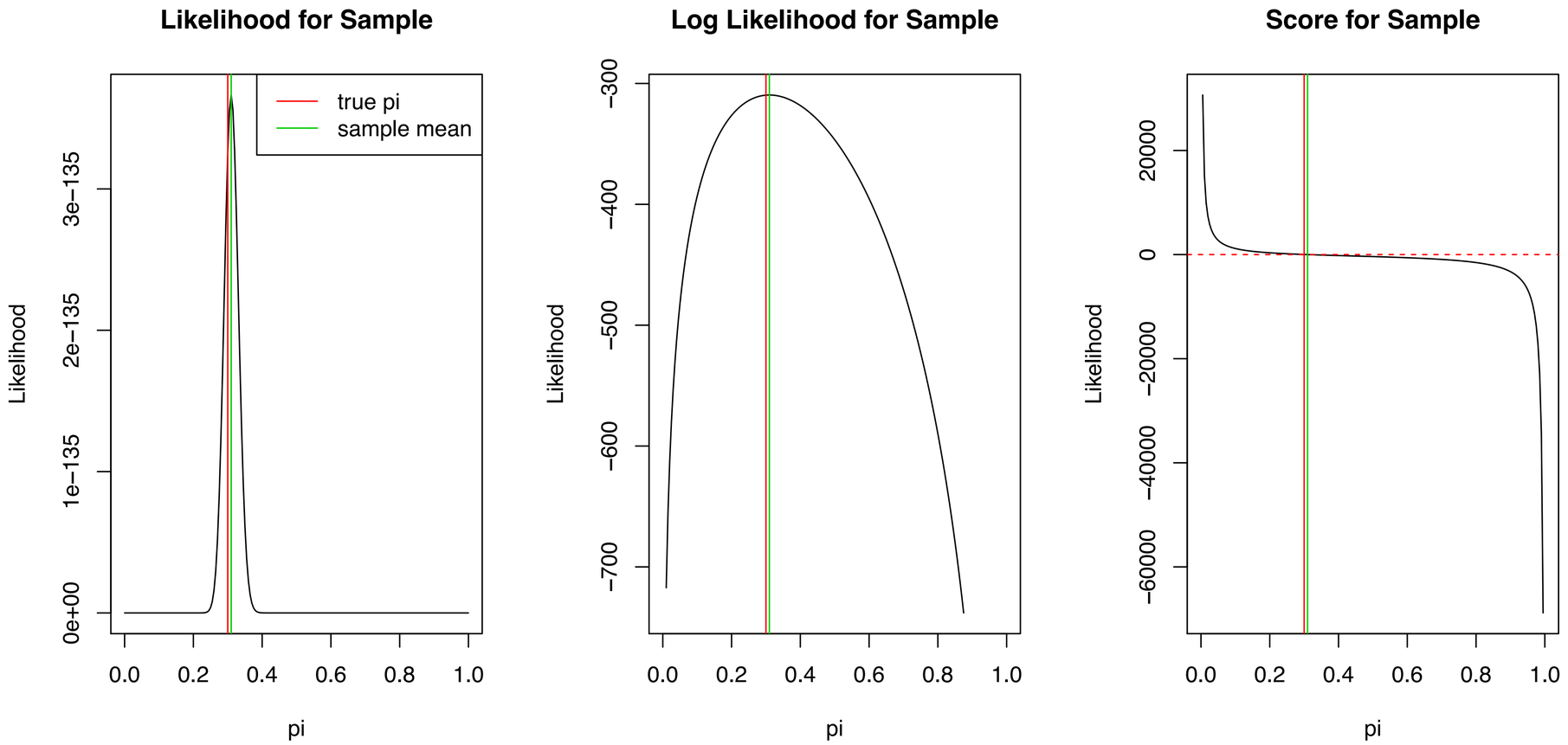
Example

With $N = 20$, $Y_i \sim Bern(\pi = 0.3)$:



Example

With $N = 500$, $Y_i \sim Bern(\pi = 0.3)$:



We *assume* data produced by a distribution, $p(\boldsymbol{y}|\theta)$.

Maximum likelihood estimation (MLE)

MLE takes the data ($\boldsymbol{y}$ here) as observed and varies θ to maximize $p(\boldsymbol{y}|\theta)$.

Notationally, we write $L(\theta|\boldsymbol{y}) = p(\boldsymbol{y}|\theta)$ as the "likelihood" at θ, and choose θ to maximize it.

We *assume* data produced by a distribution, $p(\boldsymbol{y}|\theta)$.

Maximum likelihood estimation (MLE)

MLE takes the data ($\boldsymbol{y}$ here) as observed and varies θ to maximize $p(\boldsymbol{y}|\theta)$.

Notationally, we write $L(\theta|\boldsymbol{y}) = p(\boldsymbol{y}|\theta)$ as the "likelihood" at θ, and choose θ to maximize it.

How to think of the likelihood?

We *assume* data produced by a distribution, $p(\boldsymbol{y}|\theta)$.

Maximum likelihood estimation (MLE)

MLE takes the data ($\boldsymbol{y}$ here) as observed and varies θ to maximize $p(\boldsymbol{y}|\theta)$.

Notationally, we write $L(\theta|\boldsymbol{y}) = p(\boldsymbol{y}|\theta)$ as the "likelihood" at θ, and choose θ to maximize it.

How to think of the likelihood?

- Most useful to remember it is joint density of the data, $p(\boldsymbol{y}|\theta)$.

We *assume* data produced by a distribution, $p(\boldsymbol{y}|\theta)$.

Maximum likelihood estimation (MLE)

MLE takes the data ($\boldsymbol{y}$ here) as observed and varies θ to maximize $p(\boldsymbol{y}|\theta)$.

Notationally, we write $L(\theta|\boldsymbol{y}) = p(\boldsymbol{y}|\theta)$ as the "likelihood" at θ, and choose θ to maximize it.

How to think of the likelihood?

- Most useful to remember it is joint density of the data, $p(\boldsymbol{y}|\theta)$.
- Though as argument of θ, it is not a density function

$$\int L(\theta|\boldsymbol{y})d\theta \neq 1$$

We *assume* data produced by a distribution, $p(\boldsymbol{y}|\theta)$.

Maximum likelihood estimation (MLE)

MLE takes the data ($\boldsymbol{y}$ here) as observed and varies θ to maximize $p(\boldsymbol{y}|\theta)$.

Notationally, we write $L(\theta|\boldsymbol{y}) = p(\boldsymbol{y}|\theta)$ as the "likelihood" at θ, and choose θ to maximize it.

How to think of the likelihood?

- Most useful to remember it is joint density of the data, $p(\boldsymbol{y}|\theta)$.
- Though as argument of θ, it is not a density function

$$\int L(\theta|\boldsymbol{y})d\theta \neq 1$$

- As a substitute for knowing $p(\theta|\boldsymbol{y})$, i.e. as a cheap Bayesian approach:

$$\begin{aligned} p(\theta|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\theta)p(\theta)}{p(\boldsymbol{y})} \\ &\propto p(\boldsymbol{y}|\theta)p(\theta) \end{aligned}$$

We *assume* data produced by a distribution, $p(\boldsymbol{y}|\theta)$.

Maximum likelihood estimation (MLE)

MLE takes the data ($\boldsymbol{y}$ here) as observed and varies θ to maximize $p(\boldsymbol{y}|\theta)$.

Notationally, we write $L(\theta|\boldsymbol{y}) = p(\boldsymbol{y}|\theta)$ as the "likelihood" at θ, and choose θ to maximize it.

How to think of the likelihood?

- Most useful to remember it is joint density of the data, $p(\boldsymbol{y}|\theta)$.
- Though as argument of θ, it is not a density function

$$\int L(\theta|\boldsymbol{y})d\theta \neq 1$$

- As a substitute for knowing $p(\theta|\boldsymbol{y})$, i.e. as a cheap Bayesian approach:

$$\begin{aligned} p(\theta|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\theta)p(\theta)}{p(\boldsymbol{y})} \\ &\propto p(\boldsymbol{y}|\theta)p(\theta) \end{aligned}$$

Thus $L(\theta|\boldsymbol{y})$ similar to posterior ($p(\theta|\boldsymbol{y})$), but forcing flat prior.

Suppose we don't want the same π for every i, rather:

$$Y_i \sim Bern(\pi_i)$$

Suppose we don't want the same π for every i, rather:

$$Y_i \sim Bern(\pi_i)$$

Moreover, we want π_i to be a function of X_i

Generalized linear models:

Suppose we don't want the same π for every i, rather:

$$Y_i \sim Bern(\pi_i)$$

Moreover, we want π_i to be a function of X_i

Generalized linear models:

1. A distribution (from exponential family) for each outcome

Suppose we don't want the same π for every i, rather:

$$Y_i \sim Bern(\pi_i)$$

Moreover, we want π_i to be a function of X_i

Generalized linear models:

1. A distribution (from exponential family) for each outcome
2. A linear predictor, $X_i^\top \beta$

Suppose we don't want the same π for every i, rather:

$$Y_i \sim Bern(\pi_i)$$

Moreover, we want π_i to be a function of X_i

Generalized linear models:

1. A distribution (from exponential family) for each outcome
2. A linear predictor, $X_i^\top \beta$

3a. A link function $g(\cdot)$ linking conditional mean to linear predictor,

$$g(\mathbb{E}[Y_i|X_i]) = X_i^\top \beta$$

Suppose we don't want the same π for every i, rather:

$$Y_i \sim Bern(\pi_i)$$

Moreover, we want π_i to be a function of X_i

Generalized linear models:

1. A distribution (from exponential family) for each outcome
2. A linear predictor, $X_i^\top \beta$

3a. A link function $g(\cdot)$ linking conditional mean to linear predictor,

$$g(\mathbb{E}[Y_i|X_i]) = X_i^\top \beta$$

3b. We usually think instead about the inverse link function

$$\mathbb{E}[Y_i|X_i] = g^{-1}(X_i^\top \beta)$$

Suppose we don't want the same π for every i, rather:

$$Y_i \sim Bern(\pi_i)$$

Moreover, we want π_i to be a function of X_i

Generalized linear models:

1. A distribution (from exponential family) for each outcome
2. A linear predictor, $X_i^\top \beta$

3a. A link function $g(\cdot)$ linking conditional mean to linear predictor,

$$g(\mathbb{E}[Y_i|X_i]) = X_i^\top \beta$$

3b. We usually think instead about the inverse link function

$$\mathbb{E}[Y_i|X_i] = g^{-1}(X_i^\top \beta)$$

Check: what is the link function for OLS?

Our first GLM will be the logit.

The pieces:

1. A density for $Y_i|X_i$. Here, $Y_i|X_i \sim Bern(\pi_i)$
2. Structural linear component for conditional mean, $X_i^\top \beta$
   N.B. $\mathbb{E}[Y_i|X_i] = Pr(Y_i = 1|X_i) = \pi_i$

3a. A link function connecting π_i (our $\mathbb{E}[Y_i]$) to $X_i^\top \beta$. Here:

$$g(\pi_i) = logit(\pi_i) = log\frac{\pi_i}{1-\pi_i} = X_i^\top \beta$$

3b. Equivalently, inverse link *from* the linear component,

$$\pi_i = logit^{-1}(X_i^\top \beta) = \frac{1}{1+e^{-X_i^\top \beta}}$$

Our first GLM will be the logit.

The pieces:

1. A density for $Y_i|X_i$. Here, $Y_i|X_i \sim Bern(\pi_i)$
2. Structural linear component for conditional mean, $X_i^\top \beta$
   N.B. $\mathbb{E}[Y_i|X_i] = Pr(Y_i = 1|X_i) = \pi_i$

3a. A link function connecting π_i (our $\mathbb{E}[Y_i]$) to $X_i^\top \beta$. Here:

$$g(\pi_i) = logit(\pi_i) = log\frac{\pi_i}{1-\pi_i} = X_i^\top \beta$$

3b. Equivalently, inverse link *from* the linear component,

$$\pi_i = logit^{-1}(X_i^\top \beta) = \frac{1}{1 + e^{-X_i^\top \beta}}$$

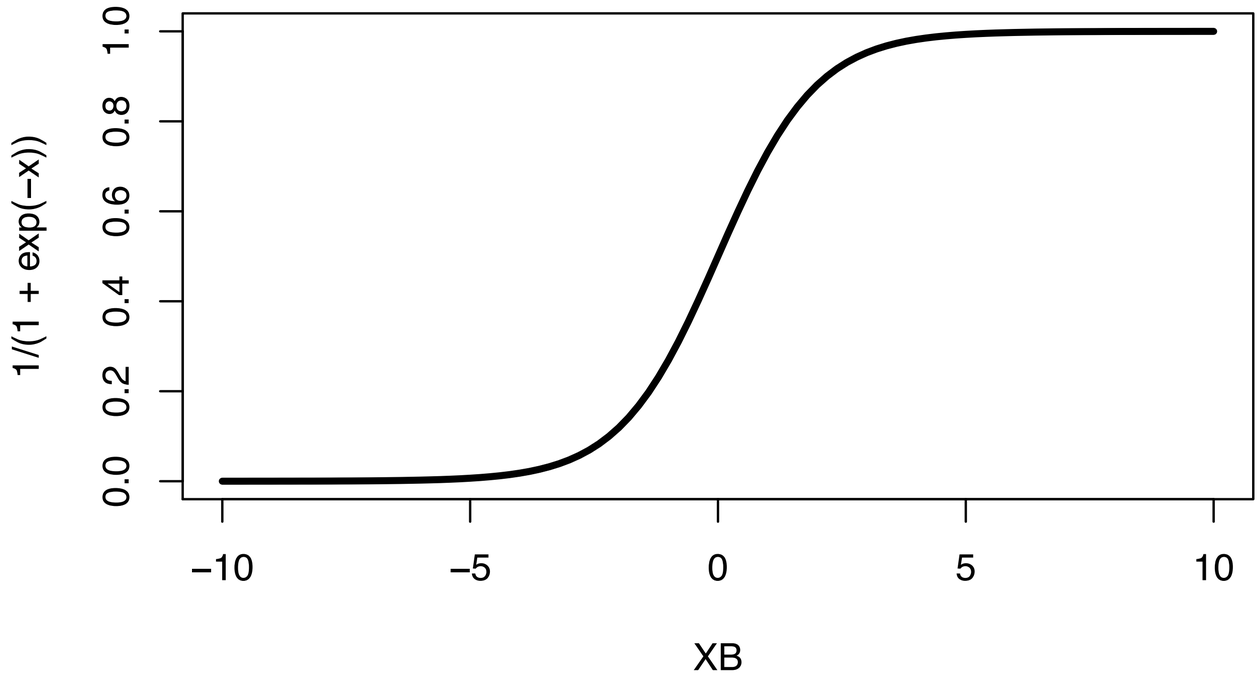
All together,

$$Y_i|X_i \sim Bern(logit^{-1}(X_i^\top \beta))$$

$X_i^\top \beta$ is in $[-\infty, \infty]$, but you need a result in $[0, 1]$.

$X_i^\top \beta$ is in $[-\infty, \infty]$, but you need a result in $[0, 1]$.

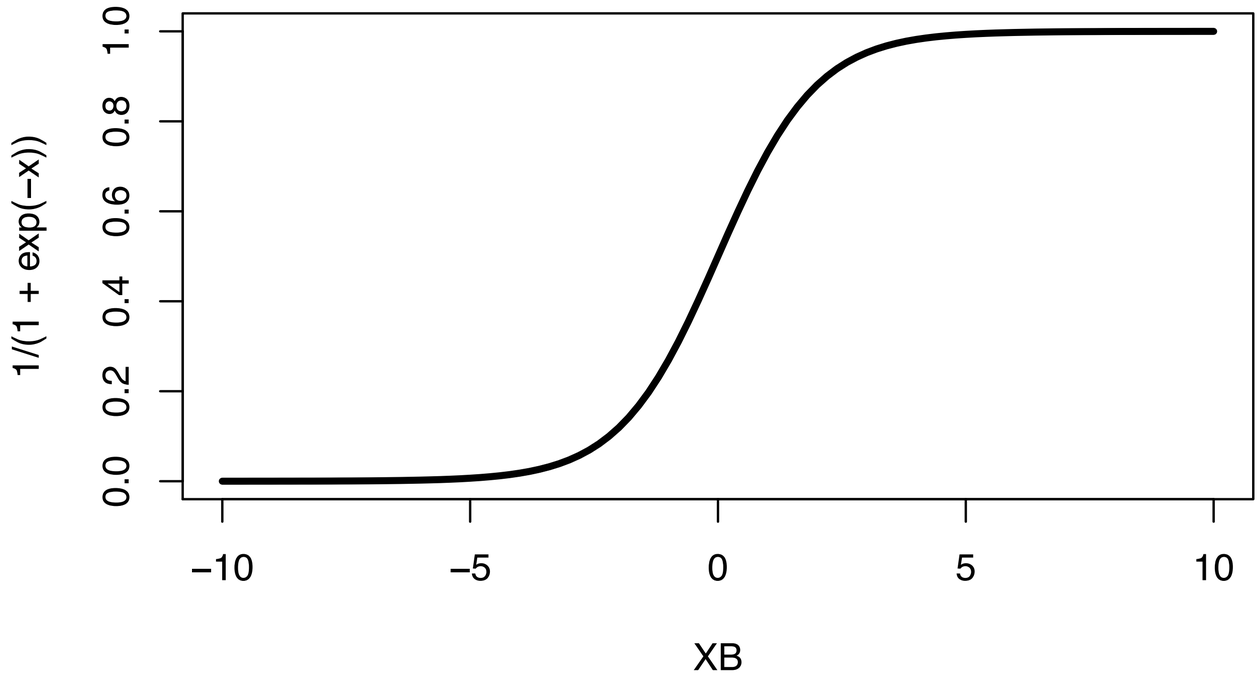
So you need a squashing function:



Why this link? No great reason, and we'll see others, but...

$X_i^\top \beta$ is in $[-\infty, \infty]$, but you need a result in $[0, 1]$.

So you need a squashing function:

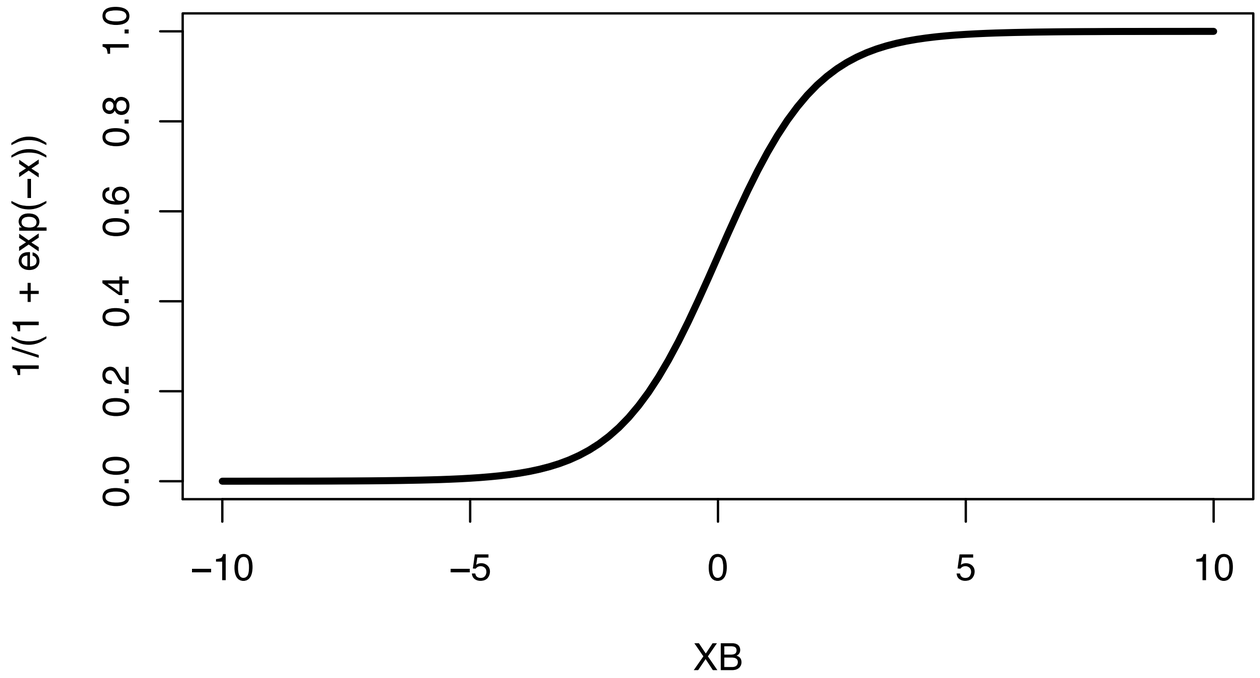


Why this link? No great reason, and we'll see others, but...

- odds: $\frac{\pi}{1-\pi} \in [0, \infty]$, so can't model that as $X_i^\top \beta$

$X_i^\top \beta$ is in $[-\infty, \infty]$, but you need a result in $[0, 1]$.

So you need a squashing function:



Why this link? No great reason, and we'll see others, but...

- odds: $\frac{\pi}{1-\pi} \in [0, \infty]$, so can't model that as $X_i^\top \beta$
- log-odds: $log(\frac{\pi}{1-\pi}) \in [-\infty, \infty]$, so you're good

Example: Civil Conflict and Political Instability

Fearon & Laitin (2003):

- war_i: Civil conflict
- $instab_i$: Political instability, 0, 1
- $lmnt_i$: Geography (log % mountainous)

Example: Civil Conflict and Political Instability

Fearon & Laitin (2003):

- war_i: Civil conflict
- $instab_i$: Political instability, 0, 1
- $lmnt_i$: Geography (log % mountainous)

We expect:

$$war_i | instab_i, lmnt_i \sim Bern(\pi_i)$$

Check: rewrite above, using logit link to relate π_i to covariates

Example: Civil Conflict and Political Instability

Fearon & Laitin (2003):

- war_i: Civil conflict
- $instab_i$: Political instability, 0, 1
- $lmnt_i$: Geography (log % mountainous)

We expect:

$$war_i | instab_i, lmnt_i \sim Bern(\pi_i)$$

Check: rewrite above, using logit link to relate π_i to covariates

Estimate the model in R

```
> glm.out=glm(war~instab+lmtnest, data=fearon_laitin, family=binomial(link="logit"))

            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.84459    0.08776  -32.41   <2e-16 ***
instab       0.90763    0.08757   10.37   <2e-16 ***
lmtnest      0.34895    0.02910   11.99   <2e-16 ***
---
```

Example: Civil Conflict and Political Instability

Fearon & Laitin (2003):

- war_i: Civil conflict
- $instab_i$: Political instability, 0, 1
- $lmnt_i$: Geography (log % mountainous)

We expect:

$$war_i | instab_i, lmnt_i \sim Bern(\pi_i)$$

Check: rewrite above, using logit link to relate π_i to covariates

Estimate the model in R

```
> glm.out=glm(war~instab+lmtnest, data=fearon_laitin, family=binomial(link="logit"))

            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.84459    0.08776  -32.41   <2e-16 ***
instab       0.90763    0.08757   10.37   <2e-16 ***
lmtnest      0.34895    0.02910   11.99   <2e-16 ***
---
```

Fitted model: $\hat{\pi}_i = \frac{1}{1+exp(2.84-.91 instab_i-.35 lmnt_i)}$

Example: Civil Conflict and Political Instability

Fearon & Laitin (2003):

- war_i: Civil conflict
- $instab_i$: Political instability, 0, 1
- $lmnt_i$: Geography (log % mountainous)

We expect:

$$war_i | instab_i, lmnt_i \sim Bern(\pi_i)$$

Check: rewrite above, using logit link to relate π_i to covariates

Estimate the model in R

```
> glm.out=glm(war~instab+lmtnest, data=fearon_laitin, family=binomial(link="logit"))

            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.84459    0.08776  -32.41   <2e-16 ***
instab       0.90763    0.08757   10.37   <2e-16 ***
lmtnest      0.34895    0.02910   11.99   <2e-16 ***
---
```

Fitted model: $\hat{\pi}_i = \frac{1}{1+exp(2.84-.91 instab_i-.35 lmnt_i)}$

So what?

How to interpret β?

- $X\beta$ is log-odds, so one-unit increase in $X^{(j)} \rightarrow \beta_j$ increase in log-odds
- usually take *exp*(β) for “odds ratio”, still not very helpful
- generally not helpful to present just a table of coefficients

How to interpret β?

- $X\beta$ is log-odds, so one-unit increase in $X^{(j)} \to \beta_j$ increase in log-odds
- usually take $exp(\beta)$ for "odds ratio", still not very helpful
- generally not helpful to present just a table of coefficients

1. Predicted probabilities when $X_i = x$:

$$\Pr(Y_i = 1 \mid X_i = x) \;=\; \pi(x) = (1 + exp(-x^\top \beta))^{-1}$$

How to interpret β?

- $X\beta$ is log-odds, so one-unit increase in $X^{(j)} \rightarrow \beta_j$ increase in log-odds
- usually take $exp(\beta)$ for "odds ratio", still not very helpful
- generally not helpful to present just a table of coefficients

1. Predicted probabilities when $X_i = x$:

$$\Pr(Y_i = 1 \mid X_i = x) \;=\; \pi(x) = (1 + exp(-x^\top \beta))^{-1}$$

2. First-Difference: average "effect" of a switch

- Suppose we want to see "effect" of instability on estimates
- $\frac{\partial \pi}{\partial instab_i}$ changes depending $instab_i$ and $lmnt_i$, so how to summarize?
- Option 1: fix $lmnt_i$ at mean (or anything) and compute:

$$\begin{aligned}
\tau &= \mathbb{E}[war_i | instab_i = 1, lmnt_i = \overline{lmnt_i}] - \mathbb{E}[war_i | instab_i = 0, lmnt_i = \overline{lmnt_i}] \\
&\approx \hat{\pi}(instab_i = 1, lmnt_i = \overline{lmnt_i}) - \hat{\pi}(war_i | instab_i = 0, lmnt_i = \overline{lmnt_i}) \\
&= logit^{-1}(\beta_0 + \beta_1 + \beta_2 \overline{lmnt_i}) - logit^{-1}(\beta_0 + \beta_2 \overline{lmnt_i})
\end{aligned}$$

How to interpret β?

- $X\beta$ is log-odds, so one-unit increase in $X^{(j)} \rightarrow \beta_j$ increase in log-odds
- usually take $exp(\beta)$ for "odds ratio", still not very helpful
- generally not helpful to present just a table of coefficients

1. Predicted probabilities when $X_i = x$:

$$\Pr(Y_i = 1 \mid X_i = x) \;=\; \pi(x) = (1 + exp(-x^\top \beta))^{-1}$$

2. First-Difference: average "effect" of a switch

- Suppose we want to see "effect" of instability on estimates
- $\frac{\partial \pi}{\partial instab_i}$ changes depending $instab_i$ and $lmnt_i$, so how to summarize?
- Option 1: fix $lmnt_i$ at mean (or anything) and compute:

$$\begin{aligned}
\tau &= \mathbb{E}[war_i | instab_i = 1, lmnt_i = \overline{lmnt_i}] - \mathbb{E}[war_i | instab_i = 0, lmnt_i = \overline{lmnt_i}] \\
&\approx \hat{\pi}(instab_i = 1, lmnt_i = \overline{lmnt_i}) - \hat{\pi}(war_i | instab_i = 0, lmnt_i = \overline{lmnt_i}) \\
&= logit^{-1}(\beta_0 + \beta_1 + \beta_2 \overline{lmnt_i}) - logit^{-1}(\beta_0 + \beta_2 \overline{lmnt_i})
\end{aligned}$$

Generally better to get average over the realized values.

See what changing your variable of interest does, over the empirical distribution of the other variables

$$
\begin{aligned}
\tau &= \mathbb{E}\left[Pr(war_i = 1 \mid instab_i = 1, lmnt_i) - Pr(war_i = 1 \mid instab_i = 0, lmnt_i)\right] \\
&= \mathbb{E}\left[\pi(instab_i = 1, lmnt_i) - \pi(instab_i = 0, lmnt_i)\right] \\
&= \int \left[\pi(1, lmnt_i) - \pi(0, lmnt_i)\right] p(lmnt_i) d(lmnt)
\end{aligned}
$$

See what changing your variable of interest does, over the empirical distribution of the other variables

$$\begin{aligned}\tau &= \mathbb{E}\left[Pr(war_i = 1 \mid instab_i = 1, lmnt_i) - Pr(war_i = 1 \mid instab_i = 0, lmnt_i)\right] \\ &= \mathbb{E}\left[\pi(instab_i = 1, lmnt_i) - \pi(instab_i = 0, lmnt_i)\right] \\ &= \int \left[\pi(1, lmnt_i) - \pi(0, lmnt_i)\right] p(lmnt_i) d(lmnt)\end{aligned}$$

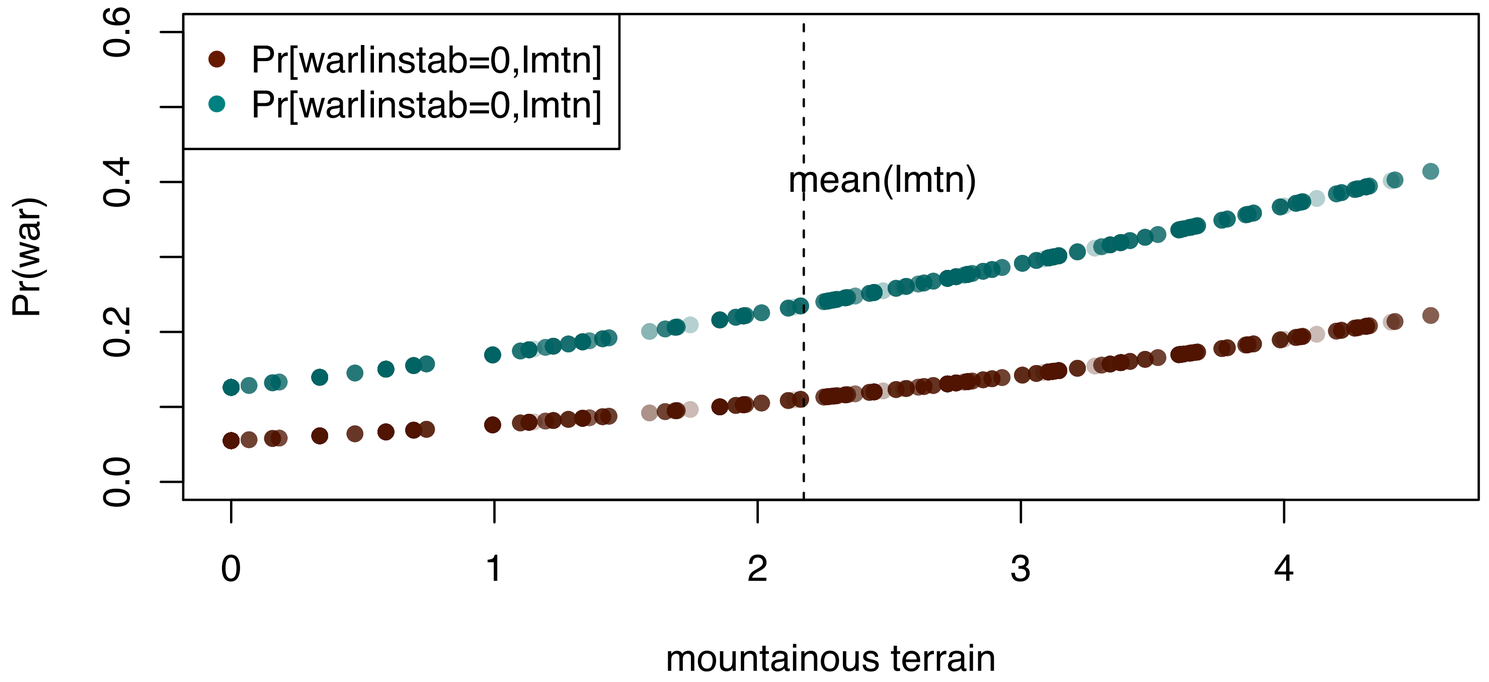
For sample analog, just make two copies of the dataset and compute:

$$\begin{aligned}\tau &= \frac{1}{N}\sum_{i=1}^{N}\{\hat{\pi}(instab = 1, lmnt_i) - \hat{\pi}(instab = 0, lmnt_i)\} \\ &= \frac{1}{N}\sum_{i=1}^{N} logit^{-1}(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 lmnt_i) - logit^{-1}(\hat{\beta}_0 + \hat{\beta}_2 lmnt_i)\end{aligned}$$

You can do same w.r.t continuous covariates using $X = x_a$ vs. $X = x_b$

```
glm.out=glm(war~instab+lmtnest, data=fearon_laitin, family=binomial(link="logit"))
beta=glm.out$coef
lmtn=fearon_laitin$lmtn

#Could use predict, or get \hat{pi} by hand for the two cases:
pi0=1/(1+exp(-beta[1]-0-beta[3]*lmtn))
pi1=1/(1+exp(-beta[1]-beta[2]-beta[3]*lmtn))
```



Quantities:

- Odds ratio $= exp(\beta_{instab}) = exp(0.91) = 2.48$
- $\hat{Pr}(war|instab = 1, \overline{lmtn}) - \hat{Pr}(war|instab = 0, \overline{lmtn}) = 0.12$
- $\frac{1}{N} \sum \left(\hat{Pr}(war|instab = 1, lmtn_i) - \hat{Pr}(war|instab = 1, lmtn_i) \right) = 0.13$

Another model for a Bernoulli outcome
Use normal CDF, $\Phi(\cdot)$

$$g(\pi_i) = \Phi^{-1}(\pi_i) \ = X_i^\top \beta$$

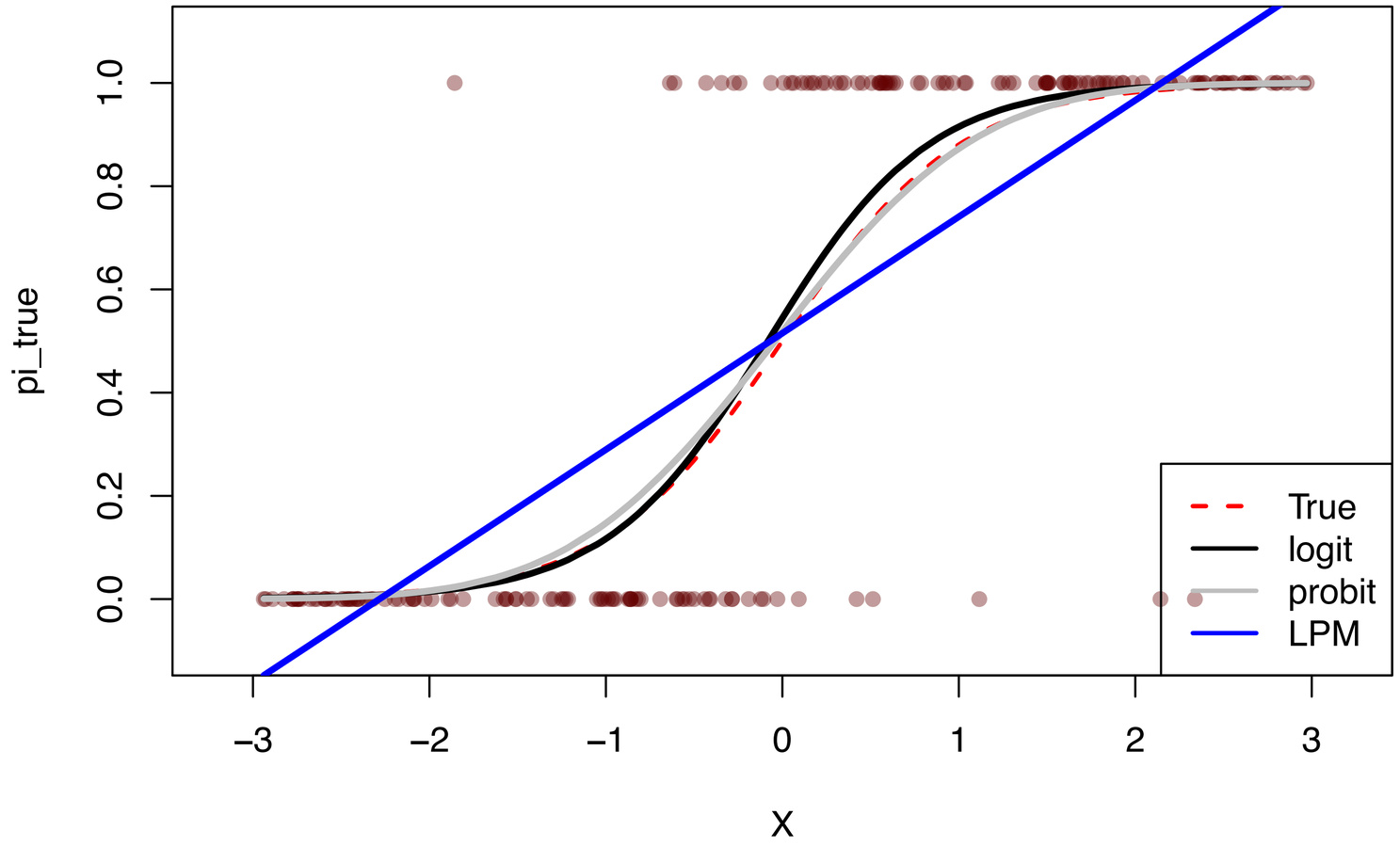
Thus, the inverse link is

$$\pi_i = \Phi(X_i^\top \beta)$$

All together,

$$Y_i | X_i \sim Bern(\Phi(X_i^\top \beta))$$

Suppose $Y_i \sim Bern(\pi_i = logit^{-1}(2X_i))$, $N = 200$



- LPM goes outside of $[0, 1]$ for extreme values of X_i
- LPM underestimates $\frac{d\pi}{dX}$ near center and overestimates near extremes
- Logit *slightly* steeper near $pi = .5$, but no practical difference

Outcome: binary

Distribution. $Y_i|X_i \sim Bernoulli(Y_i|\pi_i)$

Inverse Link. Some options:

- Logit: $\pi_i = \frac{1}{1+exp(-X_i^\top \beta)}$
- Probit: $\pi_i = \Phi(X_i^\top \beta)$
- Complementary-log-log: $\pi_i = 1 - exp(-exp(X_i^\top \beta))$

Outcome: binary

Distribution. $Y_i|X_i \sim \textit{Bernoulli}(Y_i|\pi_i)$

Inverse Link. Some options:

- Logit: $\pi_i = \frac{1}{1+exp(-X_i^\top \beta)}$
- Probit: $\pi_i = \Phi(X_i^\top \beta)$
- Complementary-log-log: $\pi_i = 1 - exp(-exp(X_i^\top \beta))$

Outcome: $\mathbb{R}$

Distribution.

- $Y_i|X_i \sim \mathcal{N}(Y_i|\theta_i)$
- $Y_i|X_i \sim t(Y_i|\theta_i)$

Inverse Link. Identity: $\mathbb{E}[Y_i|X_i] = g(X_i^\top \beta) = X_i^\top \beta$

Outcome: binary

Distribution. $Y_i|X_i \sim \textit{Bernoulli}(Y_i|\pi_i)$

Inverse Link. Some options:

- Logit: $\pi_i = \frac{1}{1+exp(-X_i^\top \beta)}$
- Probit: $\pi_i = \Phi(X_i^\top \beta)$
- Complementary-log-log: $\pi_i = 1 - exp(-exp(X_i^\top \beta))$

Outcome: $\mathbb{R}$

Distribution.

- $Y_i|X_i \sim \mathcal{N}(Y_i|\theta_i)$
- $Y_i|X_i \sim t(Y_i|\theta_i)$

Inverse Link. Identity: $\mathbb{E}[Y_i|X_i] = g(X_i^\top \beta) = X_i^\top \beta$

Outcome: count

Distribution.

- $Y_i|X_i \sim \textit{Poisson}(\lambda_i)$
- $Y_i|X_i \sim \textit{Negbin}(\mu_i, \kappa)$

Inverse Link

- For Poisson, $\mathbb{E}[Y_i|X_i] = \lambda_i = exp(X_i^\top \beta)$, (log link)
- For Negbin, $\mathbb{E}[Y_i|X_i] = \mu_i = exp(X_i^\top \beta)$, (log link)

For optimization and inference, we will often need the Score:

$$S(\theta|\boldsymbol{X}) = \frac{\partial \ell_N(\theta|\boldsymbol{X})}{\partial \theta}$$

For optimization and inference, we will often need the Score:

$$S(\theta|\boldsymbol{X}) = \frac{\partial \ell_N(\theta|\boldsymbol{X})}{\partial \theta}$$

Note,

- What is the dimensionality of $S(\theta|\boldsymbol{X})$?

For optimization and inference, we will often need the Score:

$$S(\theta|\boldsymbol{X}) = \frac{\partial \ell_N(\theta|\boldsymbol{X})}{\partial \theta}$$

Note,

- What is the dimensionality of $S(\theta|\boldsymbol{X})$? $dim(\theta)$
- We can think of score for individuals, $S_i(\theta)$, and for sample $S_N(\theta)$ (check: what is $S_i(\pi)$ for logit?)

For optimization and inference, we will often need the Score:

$$S(\theta|\boldsymbol{X}) = \frac{\partial \ell_N(\theta|\boldsymbol{X})}{\partial \theta}$$

Note,

- What is the dimensionality of $S(\theta|\boldsymbol{X})$? $dim(\theta)$
- We can think of score for individuals, $S_i(\theta)$, and for sample $S_N(\theta)$ (check: what is $S_i(\pi)$ for logit?)
- By definition, $S_N(\hat{\theta}_{MLE} \mid \boldsymbol{X}) =$?

For optimization and inference, we will often need the Score:

$$S(\theta|\boldsymbol{X}) = \frac{\partial \ell_N(\theta|\boldsymbol{X})}{\partial \theta}$$

Note,

- What is the dimensionality of $S(\theta|\boldsymbol{X})$? $dim(\theta)$
- We can think of score for individuals, $S_i(\theta)$, and for sample $S_N(\theta)$ (check: what is $S_i(\pi)$ for logit?)
- By definition, $S_N(\hat{\theta}_{MLE} \mid \boldsymbol{X}) =? 0$

Suppose $Y_i \sim \mathcal{N}(Y_i|\theta_i) = \mathcal{N}(\mu_i = X_i^\top \beta, \sigma^2)$

Note, this is equivalent to the model

$$Y_i = X_i^\top \beta + \varepsilon, \ \varepsilon \sim N(0, \sigma^2)$$

Suppose $Y_i \sim \mathcal{N}(Y_i|\theta_i) = \mathcal{N}(\mu_i = X_i^\top \beta, \sigma^2)$

Note, this is equivalent to the model

$$Y_i = X_i^\top \beta + \varepsilon, \ \varepsilon \sim N(0, \sigma^2)$$

Likelihood for one unit:

$$L_i(\beta, \sigma^2|X_i, Y_i) = p(Y_i|X_i, \beta, \sigma^2) = (2\pi\sigma^2)^{-1/2} e^{-\frac{(Y_i - X_i^\top \beta)^2}{2\sigma^2}}$$

Suppose $Y_i \sim \mathcal{N}(Y_i|\theta_i) = \mathcal{N}(\mu_i = X_i^\top \beta, \sigma^2)$

Note, this is equivalent to the model

$$Y_i = X_i^\top \beta + \varepsilon, \ \varepsilon \sim N(0, \sigma^2)$$

Likelihood for one unit:

$$L_i(\beta, \sigma^2|X_i, Y_i) = p(Y_i|X_i, \beta, \sigma^2) = (2\pi\sigma^2)^{-1/2} e^{-\frac{(Y_i - X_i^\top \beta)^2}{2\sigma^2}}$$

Likelihood for sample:

$$\begin{aligned} L_N(\theta|\boldsymbol{X}, \boldsymbol{y}) &= \prod_{i=1}^{N} (2\pi\sigma^2)^{-1/2} e^{-\frac{(Y_i - X_i^\top \beta)^2}{2\sigma^2}} \\ &= (2\pi\sigma^2)^{-N/2} e^{\frac{-1}{2\sigma^2} \sum_{i=1}^{N} (Y_i - X_i^\top \beta)^2} \end{aligned}$$

Suppose $Y_i \sim \mathcal{N}(Y_i|\theta_i) = \mathcal{N}(\mu_i = X_i^\top \beta, \sigma^2)$

Note, this is equivalent to the model

$$Y_i = X_i^\top \beta + \varepsilon, \ \varepsilon \sim N(0, \sigma^2)$$

Likelihood for one unit:

$$L_i(\beta, \sigma^2|X_i, Y_i) = p(Y_i|X_i, \beta, \sigma^2) = (2\pi\sigma^2)^{-1/2} e^{-\frac{(Y_i - X_i^\top \beta)^2}{2\sigma^2}}$$

Likelihood for sample:

$$\begin{aligned} L_N(\theta|\boldsymbol{X}, \boldsymbol{y}) &= \prod_{i=1}^{N} (2\pi\sigma^2)^{-1/2} e^{-\frac{(Y_i - X_i^\top \beta)^2}{2\sigma^2}} \\ &= (2\pi\sigma^2)^{-N/2} e^{\frac{-1}{2\sigma^2} \sum_{i=1}^{N} (Y_i - X_i^\top \beta)^2} \end{aligned}$$

Log-likelihood for the sample:

$$\ell_N(\theta|\boldsymbol{y}, \boldsymbol{X}) = -\frac{N}{2} log(2\pi) - \frac{N}{2} log(\sigma^2) - \frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\beta)^\top (\boldsymbol{y} - \boldsymbol{X}\beta)$$

$$\ell_N(\theta|\boldsymbol{y}, \boldsymbol{X}) = -\frac{N}{2}log(2\pi) - \frac{N}{2}log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)^\top(\boldsymbol{y} - \boldsymbol{X}\beta)$$

There is a score for each unit, and for sample

$$\ell_N(\theta|\boldsymbol{y}, \boldsymbol{X}) = -\frac{N}{2}log(2\pi) - \frac{N}{2}log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)^\top(\boldsymbol{y} - \boldsymbol{X}\beta)$$

There is a score for each unit, and for sample

- sum of individual scores is sample score

$$\ell_N(\theta|\boldsymbol{y}, \boldsymbol{X}) = -\frac{N}{2}log(2\pi) - \frac{N}{2}log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)^\top(\boldsymbol{y} - \boldsymbol{X}\beta)$$

There is a score for each unit, and for sample

- sum of individual scores is sample score
- we'll derive sample score directly here

$$\ell_N(\theta|\boldsymbol{y}, \boldsymbol{X}) = -\frac{N}{2}log(2\pi) - \frac{N}{2}log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)^\top(\boldsymbol{y} - \boldsymbol{X}\beta)$$

There is a score for each unit, and for sample

- sum of individual scores is sample score
- we'll derive sample score directly here

Score has two pieces: $[\frac{\partial \ell_N}{\partial \beta}, \frac{\partial \ell_N}{\partial \sigma^2}]^\top$. Set both to 0.

$$\frac{\partial \ell_N}{\partial \beta} = \frac{-1}{2\sigma^2}2\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\beta) = 0$$
$$\beta = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

$$\ell_N(\theta|\boldsymbol{y}, \boldsymbol{X}) = -\frac{N}{2}log(2\pi) - \frac{N}{2}log(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)^\top(\boldsymbol{y} - \boldsymbol{X}\beta)$$

There is a score for each unit, and for sample

- sum of individual scores is sample score
- we'll derive sample score directly here

Score has two pieces: $[\frac{\partial \ell_N}{\partial \beta}, \frac{\partial \ell_N}{\partial \sigma^2}]^\top$. Set both to 0.

$$\frac{\partial \ell_N}{\partial \beta} = \frac{-1}{2\sigma^2}2\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\beta) = 0$$
$$\beta = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

$$\frac{\partial \ell_N}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - \boldsymbol{X}\beta)^\top(\boldsymbol{y} - \boldsymbol{X}\beta) = 0$$
$$\sigma^2 = \frac{1}{N}(\boldsymbol{y} - \boldsymbol{X}\beta)^\top(\boldsymbol{y} - \boldsymbol{X}\beta)$$

Notes:

- Same result, same efficiency as OLS, but compare assumptions
- We were able to solve analytically via $S(\theta|\boldsymbol{D}) = 0$
- We can't do that with most non-linear models, but the Score and its derivative are often usful anyway...

Score

- For sample: $S(\theta|\boldsymbol{X}) = \frac{\partial \ell_N(\theta|\boldsymbol{X})}{\partial \theta}$
- For individual, $S_i(\theta|\boldsymbol{X}) = \frac{\partial \ell_i(\theta|\boldsymbol{X})}{\partial \theta}$
- Sample score is sum of individual scores

Score

- For sample: $S(\theta|\boldsymbol{X}) = \frac{\partial \ell_N(\theta|\boldsymbol{X})}{\partial \theta}$
- For individual, $S_i(\theta|\boldsymbol{X}) = \frac{\partial \ell_i(\theta|\boldsymbol{X})}{\partial \theta}$
- Sample score is sum of individual scores

The Hessian

$$H(\theta|\boldsymbol{X}) = \frac{\partial^2 \ell(\theta|\boldsymbol{x})}{\partial \theta \partial \theta^\top}$$

- Is a $dim(\theta) \times dim(\theta)$ matrix
- Again, Hessian for sample is sum of individual Hessians

Score

- For sample: $S(\theta|\boldsymbol{X}) = \frac{\partial \ell_N(\theta|\boldsymbol{X})}{\partial \theta}$
- For individual, $S_i(\theta|\boldsymbol{X}) = \frac{\partial \ell_i(\theta|\boldsymbol{X})}{\partial \theta}$
- Sample score is sum of individual scores

The Hessian

$$H(\theta|\boldsymbol{X}) = \frac{\partial^2 \ell(\theta|\boldsymbol{x})}{\partial \theta \partial \theta^\top}$$

- Is a $dim(\theta) \times dim(\theta)$ matrix
- Again, Hessian for sample is sum of individual Hessians

Preview: Information Matrix, $I(\theta|\boldsymbol{X})$

- $I(\theta|\boldsymbol{X}) = -\mathbb{E}[H(\theta|\boldsymbol{X})]$

Score

- For sample: $S(\theta|\boldsymbol{X}) = \frac{\partial \ell_N(\theta|\boldsymbol{X})}{\partial \theta}$
- For individual, $S_i(\theta|\boldsymbol{X}) = \frac{\partial \ell_i(\theta|\boldsymbol{X})}{\partial \theta}$
- Sample score is sum of individual scores

The Hessian

$$H(\theta|\boldsymbol{X}) = \frac{\partial^2 \ell(\theta|\boldsymbol{x})}{\partial \theta \partial \theta^\top}$$

- Is a $dim(\theta) \times dim(\theta)$ matrix
- Again, Hessian for sample is sum of individual Hessians

Preview: Information Matrix, $I(\theta|\boldsymbol{X})$

- $I(\theta|\boldsymbol{X}) = -\mathbb{E}[H(\theta|\boldsymbol{X})]$
- Under correct specification, also

$$I(\theta|\boldsymbol{x}) = -\mathbb{E}[H(\theta|\boldsymbol{x})] = \mathbb{E}[S(\theta|\boldsymbol{x})S(\theta|\boldsymbol{x})^\top]$$

Score

- For sample: $S(\theta|\boldsymbol{X}) = \frac{\partial \ell_N(\theta|\boldsymbol{X})}{\partial \theta}$
- For individual, $S_i(\theta|\boldsymbol{X}) = \frac{\partial \ell_i(\theta|\boldsymbol{X})}{\partial \theta}$
- Sample score is sum of individual scores

The Hessian

$$H(\theta|\boldsymbol{X}) = \frac{\partial^2 \ell(\theta|\boldsymbol{x})}{\partial \theta \partial \theta^\top}$$

- Is a $dim(\theta) \times dim(\theta)$ matrix
- Again, Hessian for sample is sum of individual Hessians

Preview: Information Matrix, $I(\theta|\boldsymbol{X})$

- $I(\theta|\boldsymbol{X}) = -\mathbb{E}[H(\theta|\boldsymbol{X})]$
- Under correct specification, also

$$I(\theta|\boldsymbol{x}) = -\mathbb{E}[H(\theta|\boldsymbol{x})] = \mathbb{E}[S(\theta|\boldsymbol{x})S(\theta|\boldsymbol{x})^\top]$$

- Especially useful because: $var(\theta_{MLE}|\boldsymbol{x}) = [I(\theta|\boldsymbol{x})]^{-1}$

For most non-linear link functions no closed-form solution to $S(\theta) = 0$

For most non-linear link functions no closed-form solution to $S(\theta) = 0$

So we use an optimization to find

$$\underset{\theta}{\textit{argmax}}\ \ell_N(\theta|\boldsymbol{X})$$

For most non-linear link functions no closed-form solution to $S(\theta) = 0$

So we use an optimization to find

$$\underset{\theta}{\textit{argmax}}\ \ell_N(\theta|\boldsymbol{X})$$

Intuition:

- You're at θ_t. Climb "uphill": $\theta_{t+1} = \theta_t + \alpha S(\theta_t)$ for some $\alpha > 0$
- Climb slower when curvature ($-H(\theta)$) is larger:

$$\theta_{t+1} = \theta_t + (-H(\theta_t)^{-1})S(\theta_t)$$

More formally: Newton-Raphson approach

More formally: Newton-Raphson approach

Approximate $S(\theta_{t+1})$ using θ_t, then set it to zero, solve for θ_{t+1}:

$$
\begin{aligned}
S(\theta_{t+1}) &= S(\theta_t) + (\theta_{t+1} - \theta_t) H(\theta_t) \\
0 &= S(\theta_t) + (\theta_{t+1} - \theta_t) H(\theta_t) \\
\theta_{t+1} - \theta_t &= -H(\theta_t)^{-1} S(\theta_t) \\
\theta_{t+1} &= \theta_t - H(\theta_t)^{-1} S(\theta_t)
\end{aligned}
$$

Setup log-likelihood function

```
#Logit log-likelihood
loglik=function(par, X, y){
  if (prod(apply(X,2,var))!=0) X=cbind(1,X)
  pi_est=1/(1+exp(-1*X%*%par))
  ll=sum(y*log(pi_est)+(1-y)*log(1-pi_est))
  return(ll)
}
```

Call it through optim

```
X=as.matrix(fearon_laitin[,c("instab","lmtnest")])
y=fearon_laitin$war

opt.out = optim(par = c(0,0,0), fn = loglik, X=X, y = y,
                  method = "BFGS", control = list(fnscale = -1),
                  hessian = TRUE)
opt.out$par
[1] -2.8445909  0.9076252  0.3489491
```

Call it through optim

```
X=as.matrix(fearon_laitin[,c("instab","lmtnest")])
y=fearon_laitin$war

opt.out = optim(par = c(0,0,0), fn = loglik, X=X, y = y,
                  method = "BFGS", control = list(fnscale = -1),
                  hessian = TRUE)
opt.out$par
[1] -2.8445909  0.9076252  0.3489491
```

Call it through optim

```
X=as.matrix(fearon_laitin[,c("instab","lmtnest")])
y=fearon_laitin$war

opt.out = optim(par = c(0,0,0), fn = loglik, X=X, y = y,
                  method = "BFGS", control = list(fnscale = -1),
                  hessian = TRUE)
opt.out$par
[1] -2.8445909  0.9076252  0.3489491
```

- BFGS chooses a quasi-Newton routine,
  - optionally accepts (analytic) gradient
  - for quadratic $\ell(\theta)$, guaranteed to find at least a local minimum
  - thus starting values can be important
- maximized by default; `fnscale=-1` flips the objective
- will return numerically estimated Hessian (`hessian=TRUE`)

Downsides:

- must assume distribution
- must assume $\mu_i = g^{-1}(X_i^\top \beta)$
- often inconsistent if assumptions fail

Downsides:

- must assume distribution
- must assume $\mu_i = g^{-1}(X_i^\top \beta)$
- often inconsistent if assumptions fail

Upside:

- unified approach to model wide variety of outcome variables
- when correct, best unbiased estimator (coming soon)
- often a good approach when you have a limited dependent variable