

Lecture 1: Likelihoods and linear models

March 27, 2018

[Go through syllabus]

How do we estimate things?

1. Specify a model
 - Function generating predictions
2. Identify plausible values for any unknown parameters
 - Maximize probability of observations given function
3. Assess uncertainty
 - Explore function around plausible values

Introduction to functions

$$\mathbf{y} = f(\mathbf{x})$$

- If \mathbf{y} is a vector, then it's a “multivariate” function
- I'll assume that \mathbf{x} is usually a vector

- We will work with derivatives:

- 1st derivatives:

$$\frac{d}{dx} f(x)$$

- 2nd derivatives:

$$\frac{d}{dx_2} \frac{d}{dx_1} f(x_1, x_2)$$

A note on notation:

- Italic: a scalar (or function)
 - Bold lowercase: a vector
 - Bold uppercase: a matrix
-
- I'll try to be clear about probabilities
 - Uppercase and not bold: random variable
 - tilde (\sim): distributions
 - e.g., $c \sim \text{Normal}(\mu, \sigma^2)$

Definitions

- Probability

- *Usage: The probability of the data given fixed values for parameters*

- $\Pr(\mathbf{y}|\boldsymbol{\theta})$
 - \mathbf{y} = data
 - $\boldsymbol{\theta}$ = parameters

- Likelihood

- *Usage: The likelihood of the parameters given fixed values of data*

- $L(\boldsymbol{\theta}; \mathbf{y})$

- Likelihood is only defined up to a constant of integration:

- $\Pr(\mathbf{y}|\boldsymbol{\theta}) = c \times L(\boldsymbol{\theta}; \mathbf{y})$

Laws of probability

1. Axiom of conditional probability

$$\Pr(X, Y) = \Pr(Y|X) \Pr(X)$$

- Often easier to specify conditional probabilities than joint probabilities

2. Definition of independent events

$$\Pr(Y) = \Pr(Y|X)$$

$$\Pr(X) = \Pr(X|Y)$$

- Necessary to simplify computation of probabilities

3. Law of total probability

$$\Pr(X) = \int \Pr(X, Y) dY$$

- Used when justifying hierarchical models

Specify a linear model

- Step 1 – Specify a linear predictor for response variable

$$y_i^* = \mathbf{x}_i \mathbf{b} = \sum_{j=1}^{n_j} x_{i,j} b_j$$

- where \mathbf{x}_i is a row of a predictor matrix \mathbf{X}
 - \mathbf{b} is a vector of parameters
-
- Step 2 – Specify a probability distribution for your response variable

$$y_i \sim \text{Normal}(y_i^*, \sigma^2)$$

-
- Maximum likelihood estimation (MLE)

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}; \mathbf{y}))$$

- Where $\hat{\boldsymbol{\theta}}$ is the MLE estimate of parameters
- Where $\operatorname{argmax}_{\boldsymbol{\theta}}(L(\boldsymbol{\theta}; \mathbf{y}))$ is the maximum value for $L(\boldsymbol{\theta}; \mathbf{y})$ that can be achieved for any value of $\boldsymbol{\theta}$
- *argmax* is done using maximization algorithms

Imagine we have two observations

- Axiom of conditional probability

$$\Pr(y_1, y_2 | \boldsymbol{\theta}) = \Pr(y_2 | y_1, \boldsymbol{\theta}) \Pr(y_1 | \boldsymbol{\theta})$$

- Usually we specify that each datum is independent

$$\Pr(y_1, y_2 | \boldsymbol{\theta}) = \Pr(y_2 | \boldsymbol{\theta}) \Pr(y_1 | \boldsymbol{\theta})$$

- Therefore:

$$L(\boldsymbol{\theta}; \mathbf{y}) = L(\boldsymbol{\theta}; y_2) L(\boldsymbol{\theta}; y_1) \propto \Pr(y_2 | \boldsymbol{\theta}) \Pr(y_1 | \boldsymbol{\theta})$$

- And taking logs

$$\log(L(\boldsymbol{\theta}; \mathbf{y})) = \log(L(\boldsymbol{\theta}; y_2)) + \log(L(\boldsymbol{\theta}; y_1))$$

And generalizing to more than two observations

$$\log(L(\boldsymbol{\theta}; \mathbf{y})) = \sum_{i=1}^{n_i} \log(L(\boldsymbol{\theta}; y_i))$$

therefore

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} \left(\sum_{i=1}^{n_i} \log(L(\boldsymbol{\theta}; y_i)) \right)$$

Why use maximum likelihood estimation?

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} (\log(L(\boldsymbol{\theta}; \mathbf{y})))$$

Where $p(\mathbf{y}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{y})$ is your specified probability distribution

1. Consistency (correct model)
2. Consistency (incorrect model)
3. Asymptotic normality

Why use maximum likelihood estimation?

1. Consistency

- Assume there's a true “data-generating process” (DGP)

$$\Pr(y_i | \boldsymbol{\theta}_{\text{true}}) \sim f(y_i | \boldsymbol{\theta}_{\text{true}})$$

- You've specified some probability model $p(y_i | \boldsymbol{\theta})$ and estimated $\boldsymbol{\theta}$ using your specified model

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} (L(\boldsymbol{\theta}; \mathbf{y}))$$

- Assume that your model “includes” the true DGP

$$f(\cdot) \in p(\cdot)$$

- Then as you collect more data

$$\text{As } n \rightarrow \infty, \text{ then } \hat{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}_{\text{true}}$$

Why use maximum likelihood estimation?

2. Consistency (incorrect model)

- Assume there's a true “data-generating process” (DGP)

$$\Pr(y_i|\boldsymbol{\theta}_{\text{true}}) \sim f(y_i|\boldsymbol{\theta}_{\text{true}})$$

- You've specified some probability model $p(y_i|\boldsymbol{\theta})$ and estimated $\boldsymbol{\theta}$ using your specified model

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} (L(\boldsymbol{\theta}; \mathbf{y}))$$

- Assume there's an optimal estimator

$$\boldsymbol{\theta}_{\text{optimal}} = \operatorname{argmin}_{\boldsymbol{\theta}} (\mathbb{E}(D_{KL}(p(y^*|\boldsymbol{\theta}) \rightarrow f(y^*|\boldsymbol{\theta}_0))))$$

where $D_{KL}(p(y^*|\boldsymbol{\theta}) \rightarrow f(y^*|\boldsymbol{\theta}_0))$ is the information lost when predicting new data y^* and approximating f as function p . This can be calculated as:

$$\boldsymbol{\theta}_{\text{optimal}} = \operatorname{argmin}_{\boldsymbol{\theta}} \left(\int \log \left(\frac{f(\boldsymbol{\theta}_{\text{true}})}{p(\boldsymbol{\theta})} \right) f(\boldsymbol{\theta}_{\text{true}}) dy_i \right)$$

- Then as you collect more data

$$\text{As } n \rightarrow \infty, \text{ then } \hat{\boldsymbol{\theta}} \rightarrow \boldsymbol{\theta}_{\text{optimal}}$$

Why use maximum likelihood estimation?

3. Asymptotic normality

- Assume there's an optimal estimator

$$\boldsymbol{\theta}_{optimal} = \operatorname{argmin}_{\boldsymbol{\theta}} \left(\int \log \left(\frac{f(\boldsymbol{\theta}_{true})}{p(\boldsymbol{\theta})} \right) f(\boldsymbol{\theta}_{true}) d\mathbf{y}_i \right)$$

- You've specified some probability model $p(y_i|\boldsymbol{\theta})$ and estimated $\boldsymbol{\theta}$ using your specified model

$$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} (L(\boldsymbol{\theta}; \mathbf{y}))$$

- As sample sizes get big ($n \rightarrow \infty$), if you replicate an estimator:

$$\hat{\boldsymbol{\theta}} \sim \text{MVN}(\boldsymbol{\theta}_{optimal}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ decreases with increasing n

Implications

- If you have a simulation design...
 - ... and the model used to simulate data is identical to the model used to estimate parameters
 - Estimated parameters will be perfect with large sample sizes
 - Total error will go to zero with large sample sizes
 - ... and your estimation model doesn't match the simulation model
 - Estimated parameters will converge on values with large sample sizes
 - Total error will decrease to an asymptote

Example #1 – What is the mean density of canary rockfish in the California Current?

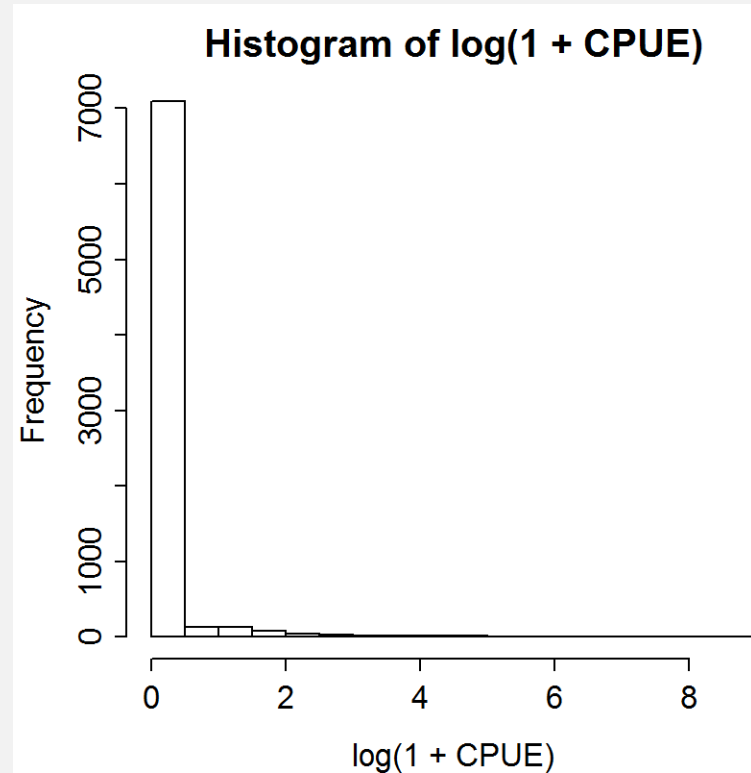
- Define linear predictor matrix

$$x_i = 1$$

– i.e.,

$$\mathbf{X} = \mathbf{1}$$

– We call \mathbf{X} an intercept matrix



-
- Approach 1 – Use existing R functions
 - Step 1 – Find function
 - For linear model, use *lm* in the base package
 - Step 2 – Apply function
 - Usually easy in R
 - Step 3 – Extract information from object
 - Often hard
 - Sometimes use *summary* or *attributes* commands

-
- Approach 1
 - [See R code]

-
- Approach 2 – Build your own code
 - Step 1 – make function for log-likelihood
 - Step 2 – use nonlinear minimizer to find maximum likelihood estimate
 - Step 3 – estimate standard errors

- How to estimate standard errors?

- Estimate the “Hessian” at the MLE

$$H(\boldsymbol{\theta}; \mathbf{y}) = \begin{bmatrix} \frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_1^2} & \frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_1 \partial \theta_2} & \frac{\partial^2 \ln L(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_2^2} \end{bmatrix}$$

- Calculate its inverse

$$\widehat{Var}(\boldsymbol{\theta}; \mathbf{y}) = \mathbf{H}^{-1}$$

- Extract element and take square root

$$\widehat{SE}(\theta_1; \mathbf{y}) = \sqrt{\widehat{Var}(\boldsymbol{\theta}; \mathbf{y})_{1,1}}$$

-
- Approach 2
 - [See R code]

- Approach 3 – Use TMB

- Step 1 – Define TMB template file

- Uses C++ code

- Step 2 – Define inputs for TMB

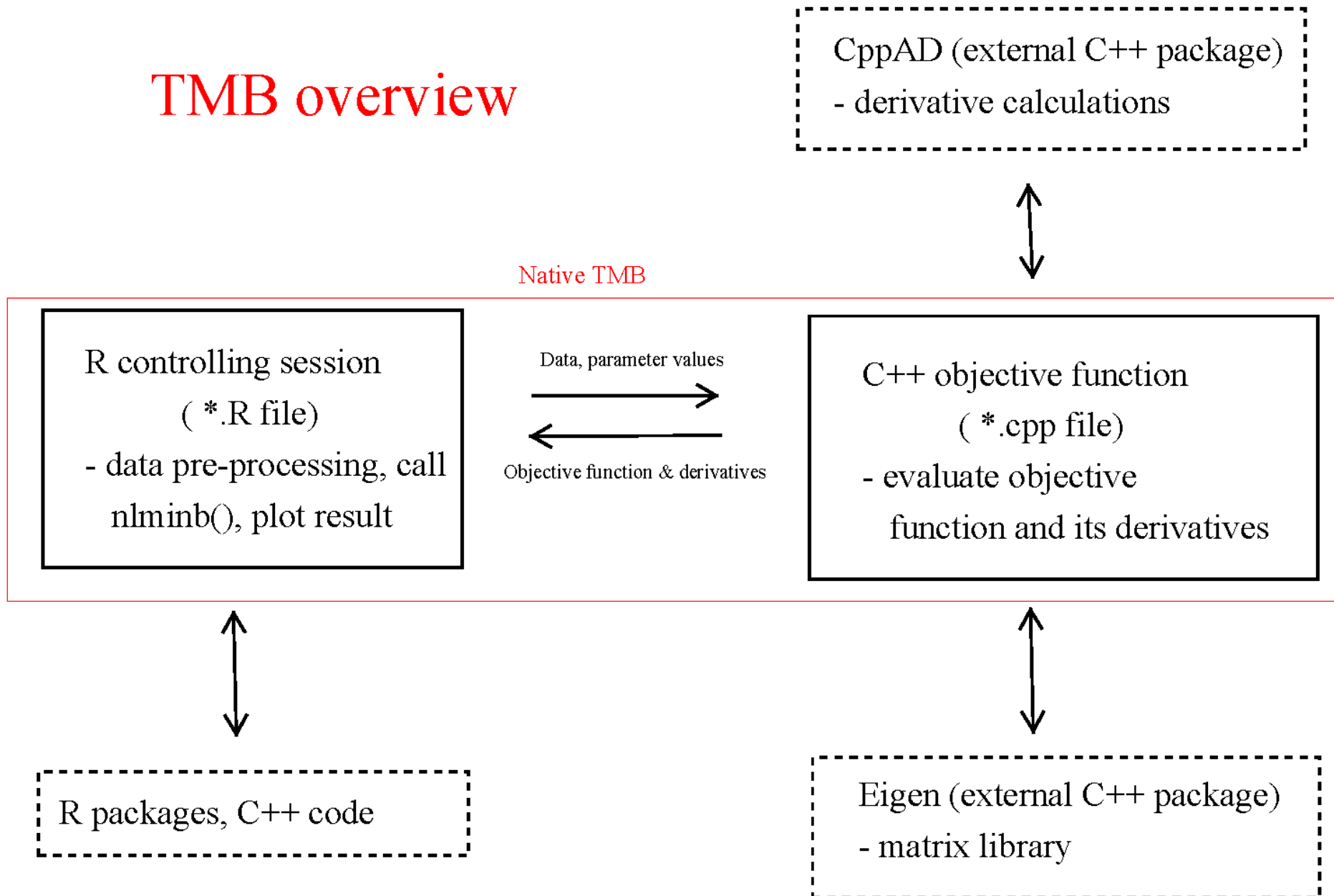
- List of “tagged” (named) elements for data and starting parameters

- Step 3 – Run optimizer in R

- Nonlinear optimizers using gradients

- Step 4 – Check model diagnostics

TMB overview



-
- Approach 3
 - [See R code]

How to know you understand a model?

1. Make predictions about behavior, and double check predictions
2. Simulation experiments

Next step:

- Add covariates (pass and latitude)
- Prediction: Adding fixed effects will always decrease the residual variance in a linear model

Testing prediction: effect of adding linear predictors

- [See R code]
- Was the prediction supported?