Topic 1: Linear regression with regularization and noise addition

Eric B. Laber

Department of Statistical Science, Duke University

Statistics 561

Perhaps, if I am very lucky, the feeble efforts of my lifetime will someday be noticed, and maybe, in some small way, they will be acknowledged as the greatest works of genius ever created by Man.
—Eric Laber

Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write!
—Mark Sargent

Warm-up (5 minutes)

- Explain to your group
  - What are the three major areas of machine learning?
  - What is the bias-variance trade-off?
  - What's the difference between stats/ML and why is this course called probabilistic ML?[1]
- True or false
  - The R^2 of regressing Y on X_1 and X_2 is always no greater than the sum of the R^2s of Y on X_1 and Y on X_2
  - $\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}^\intercal \beta^*$ then $Y = \boldsymbol{X}^\intercal \beta^* + \epsilon$, w/ $\mathbb{E}(\epsilon|\boldsymbol{X}) = 0$ a.s.
  - The game of cheese-rolling was invented in Cheddarshire, England by pagans during Roman Occupation

[1] Or whatever it's called.

Roadmap

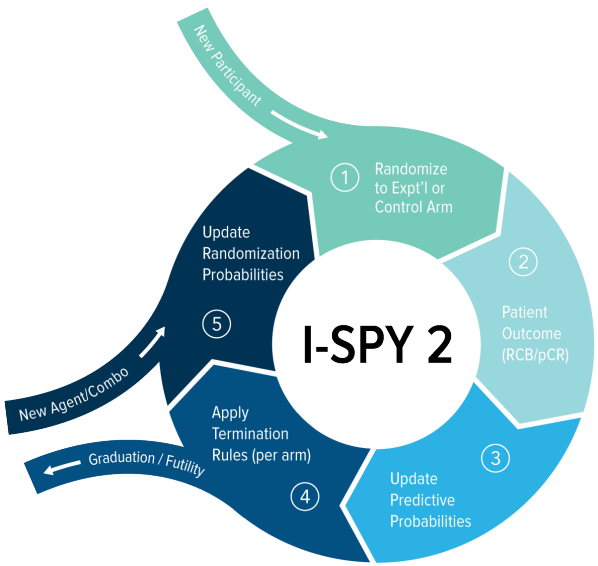
- Background and biases
- Course objectives (i.e., what's not here)
- Notation
- Linear model basics

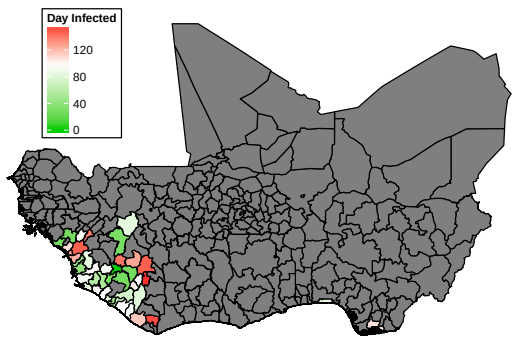
Roadmap

- **Background and biases**
- Course objectives (i.e., what's not here)
- Notation
- Linear model basics

Introduction, background, and biases

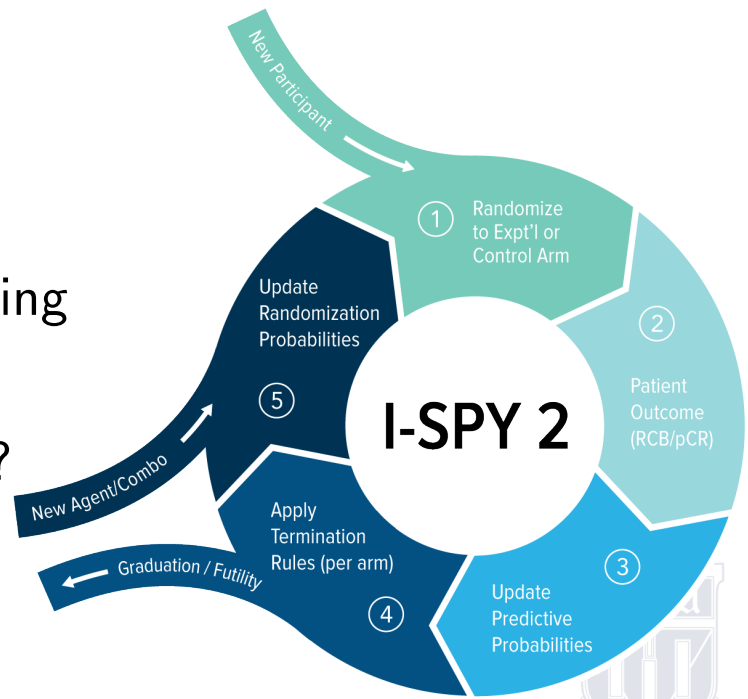
- Research focuses on methods development for optimal sequential decision making in data-impoverished settings
  - Applications in precision medicine, public health, and logistics
  - Even if data volume high, signal is weak
  - Statistical inference is critical in these areas





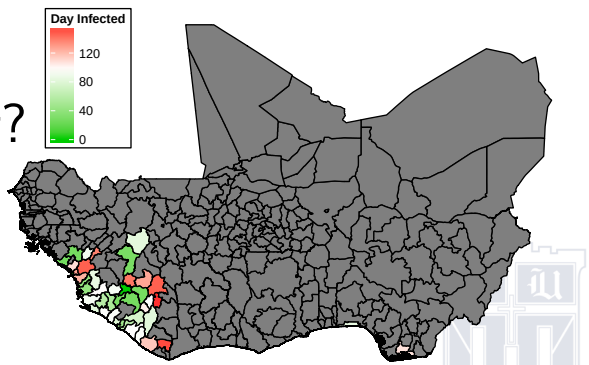
Ex. ISPY 2+

- Adaptive platform clinical trial for breast cancer
- Treatment randomization probabilities based on accumulating info on treatment efficacy/side-effects
- Goal: optimize health outcomes for current (in-trial) and future patients
- Key questions:
  - How should treatment be tailored to evolving patient characteristics?
  - When, if, for whom does a given txt work?
  - What patient characteristics (biomarkers) are critical for choosing treatment?



Ex. Ebola Virus Disease

- Allocation of interventions over space and time
- Logistical and cost constraints + imperfect information
- Goal: min mortality and morbidity with limited resources
- Key questions:
  - How should interventions be tailored to evolving pandemic characteristics?
  - Is RL-based intervention strategy better than an expert strategy? If so, how much better? Why is it better?



Ex. Bookstore assortment selection

- Which titles should be offered at Amazon Books?
- Enormous actions space, e.g., 100K choose 2K
- Goal: delight customers and optimize profit
- Key questions:
  - What is the optimal assortment?
    How does it change over time/space?
  - How does RL-based assortment differ
    from expert curators?
  - How to experiment efficiently and safely?

Ex. Bookstore location selection

- Where to place physical stores?
- Each action associated with large investment
- Goal: delight customers and optimize profit
- Key questions:
  - Which subset of available sites should be chosen?
  - What makes a 'good' location?
  - How to optimally grow portfolio?

Informing human decisions with RL

- ML/RL is now being given serious consideration in high-stakes problems previously restricted to human decision makers
  - Precision medicine
  - Public health
  - Low-frequency business decisions
  - Defense and intelligence
  - ...
- Transitioning to ML/RL-based decision systems requires trust developed by knowledge generation and uncertainty quantification $\Rightarrow$ statistical inference is critical

ML/RL in practice

- How to use data to make a good decision?
  - I.e., set the price of a product, assign a treatment, run a promotion, adjust trajectory of autonomous vehicle, etc.
  - Predict what will happen using Stats/ML model
  - Quantify uncertainty/risk/DSI[2]
  - Select decision based on longterm criterion

[2] Downstream impacts

ML/RL in practice

- How to use data to make a good decision?
  - I.e., set the price of a product, assign a treatment, run a promotion, adjust trajectory of autonomous vehicle, etc.
  - Predict what will happen using Stats/ML model
  - Quantify uncertainty/risk/DSI[2]
  - Select decision based on longterm criterion
- What happens when we make a decision?
  - Generate utility
  - Generate information
  - Set ourselves up for subsequent utility

[2] Downstream impacts

Expert decision making

I'm like really good at business. I've got a brief case and everything. –Juliet Laber, Age 4

ML/RL + expert decision making

- Old model: ask nerdery for forecast then 'go do business'
  - Siloed workflow with design/experimentation, decision making, evaluation weakly integrated (if at all)
  - Unchecked/unjustified human overrides rampant
- Emerging model: build end-to-end decision support systems that drive decisions including if, how, and when to experiment
  - Off-the-shelf methods are almost never satisfactory
    - Complex sampling schemes
    - Partial control over system
    - Non-stationary environment
    - Cost and/or logistic constraints
    - Data quality
    - ...
  - Bespoke but rigorous methods needed

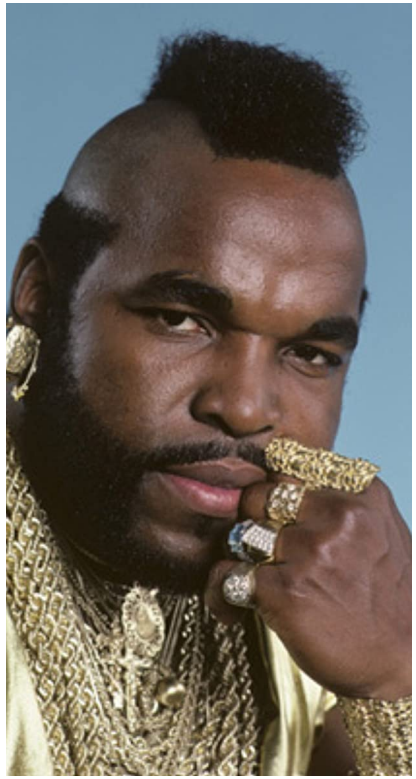
Required BS on data science

- Lots of propaganda about data science unicorns
  - Software development engineering
  - Statistics/ML experts
  - Data(base) engineering
  - Domain experts
- Trying to be an expert in all of these is pointless. Anyone who claims they are is either delusional or has no understanding of what deep expertise looks like.[3]

[3] A third option is that the person you met is Craig Citro. Who, somehow, is awesome at all these things. So, if you're Craig Citro, you can ignore.

Required BS on data science cont'd

- Be the 'T'
  - Deep in one area broad and others
  - This course could be the start of going deep in ML/RL or broadening your knowledge as you specialize elsewhere
- Data science v stats is not flexible vs parametric models
  - sklearn and keras $\neq$ data science
  - An effective data scientist has strong intuition for statistics and computing and the depth/technical expertise to solve non-standard problems



Roadmap

- Background and biases
- Course objectives (i.e., what's not here)
- Notation
- Linear model basics

Roadmap

- Background and biases
- **Course objectives (i.e., what's not here)**
- Notation
- Linear model basics

What's this course about?

- Introduce the foundations of estimation and inference in prediction and decision problems
  - Confidence intervals and prediction intervals
  - Asymptotic and (some) non-asymptotic techniques
- Overarching goal of this course is to prepare you to digest the broader ML/RL literature and develop your own methods

How were topics selected?

- Sought simplest problems/ideas/techniques that illustrate salient features of what we want to learn
  - Heavy focus on linear models
  - Kernel methods for nonlinear models
  - Mostly regression and decision making but some on classification (and maybe dim reduction)
  - Homework may involve asking you to learn/apply methods not covered in class

What's not in this course?

- This is not a catalog of all ML/RL methods
  - Such a course is impossible w/ limited time and it would be immediate out-of-date as new methods arise all the time
  - YouTube, coursera, etc. do an excellent job of giving brief reviews of new methods along with use cases etc.[4]
- Neural networks
  - WHAT?!?!!?!!
  - Time-permitting, I may talk about this when in our section in RL, you can also use Deep NNets in your projects
  - If there's enough pressure I may do some extra lectures on this, aw shucks, I just want you to like me

[4] This is awesome btw, and I strongly encourage you to make use of the abundance of free and high-quality content.

Grading

- Grades will be assigned according to
  - Homework 50% (approx 12 assignments, drop 2 lowest)
  - Exams 30% (two, take-home)
  - Project 20% (group, do something awesome)
- I don't have a problem giving all A's
  - Work with your classmates no reason to do poorly on HW's
  - I want you to learn. I'd rather go deep on a handful of topics than create superficial understanding on many topics. Due to COVID, we'll need to be flexible and adaptable.

Lectures

- Delivered live but recordings will be posted
- Mix of derivations/proofs and slides
- Some coding in but mostly in TA sessions
  - Code should be submitted as jupyter notebook
  - Readability counts
- Do the in-class activities!

Office hours

- Right after class 10-11AM (zoom)
- Any extra review sessions etc. will probably take place Sunday nights (to avoid conflicts with other courses etc.)
- Your TAs will set their own OHs

Questions?

- Questions? Comments? Concerns?

Roadmap

- Background and biases
- Course objectives (i.e., what's not here)
- Notation
- Linear model basics

Roadmap

- Background and biases
- Course objectives (i.e., what's not here)
- **Notation**
- Linear model basics

Let's get into it (finally)

- Upper case letters, e.g., $(\boldsymbol{X}, Y)$ for random variables and lowercase letters, e.g., $(\boldsymbol{x}, y)$, for their instances
- Generally bold vectors and matrices, e.g., $\boldsymbol{X}$ and $\boldsymbol{\Omega}$
- Given $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim_{i.i.d.} P$ we use operator notation to denote expectation and average

$$
\begin{aligned}
Pf(\boldsymbol{X}) &= \int f(\boldsymbol{x}) dP(\boldsymbol{x}) \\
\mathbb{P}_n f(\boldsymbol{X}) &= \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i),
\end{aligned}
$$

where $\mathbb{P}_n$ is called empirical measure

Let's practice

- Suppose $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^p$ are ind copies of $\boldsymbol{X} \sim P$
- The strong law of large numbers says that if $\mu = P\boldsymbol{X}$ exists then $n^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i \to \mu$ almost surely, using operator notation, we have

$$(\mathbb{P}_n - P)\boldsymbol{X} \to 0 \text{ a.s.}$$

- Notation is more compact but can take some getting used to

More on operator notation

- Suppose we observe $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n) \sim_{i.i.d.} P$ where $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \mathbb{R}$
- Construct $\widehat{f}_n : \mathbb{R}^p \to \mathbb{R}$ from data then

$$P\widehat{f}_n(\boldsymbol{X}) = \int \widehat{f}_n(\boldsymbol{x}) dP(\boldsymbol{x})$$

averages only over distn of $\boldsymbol{X}$, thus is a random quantity because it depends on the data

  - Write $\mathbb{E}\widehat{f}_n(\boldsymbol{X})$ to denote average over distn of $\boldsymbol{X}$ *and* the observed data
  - Useful for evaluating performance of est predictive model

More practice

- Assume $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \sim_{i.i.d.} P$ with finite mean and invertible covariance matrix $\boldsymbol{\Sigma}$, state the central limit theorem (CLT) using operator notation

Roadmap

- Background and biases
- Course objectives (i.e., what's not here)
- Notation
- Linear model basics

Roadmap

- Background and biases
- Course objectives (i.e., what's not here)
- Notation
- **Linear model basics**

Setup

- Observe $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ comprising n ind copies of $(\boldsymbol{X}, Y) \sim P$
  - $\boldsymbol{X} \in \mathbb{R}^p$: vector of inputs (predictors)
  - $Y \in \mathbb{R}$: scalar output (response)
- Goal: predict output at new vector of inputs

Shaking off the rust

- Let $f : \mathbb{R}^p \to \mathbb{R}$ be a map from inputs to outputs
- Define $\mu(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$, then the expected squared error of f at a new input-output pair

$$\begin{aligned} P\{Y - f(\boldsymbol{X})\}^2 &= P\{Y - \mu(\boldsymbol{X}) + \mu(\boldsymbol{X}) - f(\boldsymbol{X})\}^2 \\ &= P\{Y - \mu(\boldsymbol{X})\}^2 + \mathbb{E}\{\mu(\boldsymbol{X}) - f(\boldsymbol{X})\}^2 \end{aligned}$$

  which is minimized when $f(\boldsymbol{x}) = \mu(\boldsymbol{X})$ for all $\boldsymbol{x}$
- Least squares viewed as estimating conditional mean

Linear models

- Posit a linear working model $\mu(\boldsymbol{x};\boldsymbol{\beta}) = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{\beta}$ for $\mu(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$, which we estimate using least squares

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_n &= \arg\min_{\boldsymbol{\beta}} \mathbb{P}_n \{Y - \mu(\boldsymbol{X};\boldsymbol{\beta})\}^2 \\ &= \{\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}\}^{-1} \mathbb{P}_n \boldsymbol{X} Y,\end{aligned}$$

  where we've assumed $\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}$ is invertible

- Note that $\widehat{\boldsymbol{\beta}}_n$ is also a root (in $\boldsymbol{\beta}$) of

$$\mathbb{P}_n(Y - \boldsymbol{X}^{\mathsf{T}}\boldsymbol{\beta})\boldsymbol{X} = 0$$

Left blank for notes

Characterizing behavior of $\widehat{\beta}_n$

- Our v. simple perspective on estimation and inference
  - Construct estimator
  - Study properties of that estimator
- Fundamental idea: sampling distn of a estimator
  - $\widehat{\beta}_n$ is constructed from $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ and thus it has a distn induced by these random variables
  - The term sampling distn reflects the fact that the distn is induced by repeated sampling data sets of size n using the same mechanism by which the original sample was constructed

Where does $\widehat{\beta}_n$ concentrate?

- We do not assume the linear model is correctly specified
- Define the population analog of $\widehat{\beta}_n$ as

$$\begin{aligned} \beta^* &= \arg\min_{\beta} P\left\{Y - \mu(\boldsymbol{X};\beta)\right\}^2 \\ &= \left\{P\boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\}^{-1} P\boldsymbol{X}Y, \end{aligned}$$

  where we've assumed $P\boldsymbol{X}\boldsymbol{X}^\mathsf{T}$ is invertible

- Note that β^* is a root (in β) of

$$P(Y - \boldsymbol{X}^\mathsf{T}\beta)\boldsymbol{X} = 0$$

Review: consistency of least squares estimator

- Claim: if all requisite moments exist, then $\widehat{\beta}_n \to \beta^*$ a.s.

Review: asymptotic normality of least squares estimator

- Claim: if all requisite moments exist, then $\sqrt{n}(\widehat{\beta}_n - \beta^*)$ is asymptotically normal with mean zero and covariance $\mathbf{\Sigma}$

Left blank for notes

Confidence regions for β^*

- $(1-\alpha) \times 100\%$ Wald-type confidence set for β^*

$$\zeta_{1-\alpha,n} = \left\{ \beta \in \mathbb{R}^p \, : \, n(\widehat{\beta}_n - \beta)^\intercal \widehat{\Sigma}_n^{-1} (\widehat{\beta}_n - \beta) \leq \chi^2_{1-\alpha,p}, \right\}$$

  where $\chi^2_{1-\alpha,p}$ is the $(1-\alpha) \times 100$ percentile of χ^2-distn with p degrees of freedom

- Why does the Wald-type confidence set work?

Confidence intervals (CIs) for $x^\intercal \beta^*$

- Observe new feature $X = x$ report $x^\intercal \widehat{\beta}_n$ as point estimate of $x^\intercal \beta^*$, want to attach measure of uncertainty
  - Construct CI for $x^\intercal \beta^*$ using asy normality
  - We've shown $\sqrt{n}(\widehat{\beta}_n - \beta^*) \rightsquigarrow N(0, \Sigma)$ thus

$$x^\intercal \sqrt{n} \left(\widehat{\beta}_n - \beta^* \right) \rightsquigarrow N\left(0, x^\intercal \Sigma x\right),$$

  so that as $(1-\alpha) \times 100\%$ asymptotic CI for $x^\intercal \beta^*$ is $x^\intercal \widehat{\beta}_n \pm z_{1-\alpha/2} \sqrt{x^\intercal \widehat{\Sigma}_n x / n}$

How good is our predictive model?

- Measures of error
  - Residual error $R \triangleq P\left(Y - \boldsymbol{X}^{\mathsf{T}}\beta^*\right)^2$
  - Conditional error $C(\widehat{\beta}_n) \triangleq P\left(Y - \boldsymbol{X}^{\mathsf{T}}\widehat{\beta}_n\right)^2$
  - Average conditional error $A_n = \mathbb{E}C(\widehat{\beta}_n)$
- Which measure is appropriate depends on the scientific question of interest[5]

[5] When we discuss decision problems we'll see that these notions of error need not agree even in infinite samples. This is somewhat counterintuitive and has caused confusion in the literature.

Residual error

- How good is a linear model in this context?
- Deriving a confidence interval for R is straightforward

$$\sqrt{n}\left\{\mathbb{P}_n\left(Y - \boldsymbol{X}^\intercal \widehat{\beta}_n\right)^2 - R\right\} \quad \rightsquigarrow \quad ???$$

Left blank for notes

Conditional error

- How good is the model I fit?
- Note that the conditional error is a random quantity, i.e., a data-dependent estimand, which is somewhat unusual
- Derive limiting distribution of

$$\sqrt{n}\left(\mathbb{P}_n - P\right)\left(Y - \boldsymbol{X}^{\mathsf{T}}\widehat{\beta}_n\right)^2$$

Left blank for notes

Discussion of conditional error

- Want error given estimated model
- Might want conditional confidence interval, e.g., a random set $\widehat{I}_n$ such that

$$P\left\{P(Y - \boldsymbol{X}^\intercal \widehat{\beta}_n)^2 \in \widehat{I}_n \middle| \widehat{\beta}_n\right\} \geq 1 - \alpha + o_P(1)$$

- Constructing conditional confidence sets is generally more difficult/complex than unconditional sets

Deriving a conditional error (handwaving)

Average conditional error

- Useful for comparing algorithm in a given domain
- Depends on training set size (but not random)
- What will happen if we try an use an asymptotic approximation (i.e., one where $n \to \infty$)
- How does $A_n = \mathbb{E}C(\widehat{\beta}_n)$ decay with n?

Derive approximation for A_n

[Hint: use normality of $\widehat{\beta}_n$]

Lack of linearity

- We have not assumed the linear model is correctly specified
  - Aligned with perspective in this class: define estimand, posit model, construct estimator, study properties of the estimator
  - Introduce assumptions only if/when needed, this helps us gain clarity on nature or problem and properties of estimator
- Meaning and misspecification
  - Inference for projection of $\mathbb{E}(Y|\boldsymbol{X})$ on linspan of $\boldsymbol{X}$
  - Inference for measures of performance meaningful even when model misspecified, i.e., we can always ask how well the model we fit will perform if used to make predictions

Further complications with misspecification: covariate shift

- Suppose we have two distributions P and P' for $(\boldsymbol{X}, Y)$ s.t.
  - Conditional distribution of $Y|X$ is the same under P and P'
  - Distribution of $\boldsymbol{X}$ different under P and P'
- Projection of $\mathbb{E}(Y|\boldsymbol{X})$ onto linspan of $\boldsymbol{X}$ is
  - $\beta^*_P = (P\boldsymbol{X}\boldsymbol{X}^\intercal)^{-1} P\boldsymbol{X}Y$ under P
  - $\beta^*_{P'} = (P'\boldsymbol{X}\boldsymbol{X}^\intercal)^{-1} P'\boldsymbol{X}Y$ under P'

  thus even though the conditional mean $\mathbb{E}(Y|\boldsymbol{X})$ is the same under P and P' the projections are different

Further complications with misspecification: covariate shift cont'd

- Show that if $\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = \boldsymbol{x}^\intercal \beta^*$ then $\beta^*_P = \beta^*_{P'}$

Discussion

- Whirlwind tour of lin models in regression
  - Omitted diagnostics (plots, GOF)
  - Omitted theory for correctly specified model[6]
- In HW 1 you'll deep dive into some of the key ideas presented here the ideas presented here
- Next time: regularization

[6] This is a special case of what we covered here but one obtains several nice properties and simplifications if the model is correct. For review of linear models, Seber and Lee (2003) is an excellent reference.

Thank you.

eric.laber@duke.edu

laber-labs.com