Soc504: Regularization and Hierarchical Models[1]

Brandon Stewart

Princeton

April 24-26, 2017

[1] I am grateful to Justin Grimmer, Marc Ratkovic and Dustin Tingley for sharing their slides with me. Some figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani."

Housekeeping

Housekeeping

- Replication papers: formats, posters, deadlines

Housekeeping

- Replication papers: formats, posters, deadlines
- Poster session timing

Housekeeping

- Replication papers: formats, posters, deadlines
- Poster session timing
- Schedule for the final week of class

Housekeeping

- Replication papers: formats, posters, deadlines
- Poster session timing
- Schedule for the final week of class
- After the final week, feedback etc.

Housekeeping

- Replication papers: formats, posters, deadlines
- Poster session timing
- Schedule for the final week of class
- After the final week, feedback etc.
- Some notes on lecture structure for this week.

Readings

- Murphy (2012) *Machine Learning: a Probabilistic Perspective*

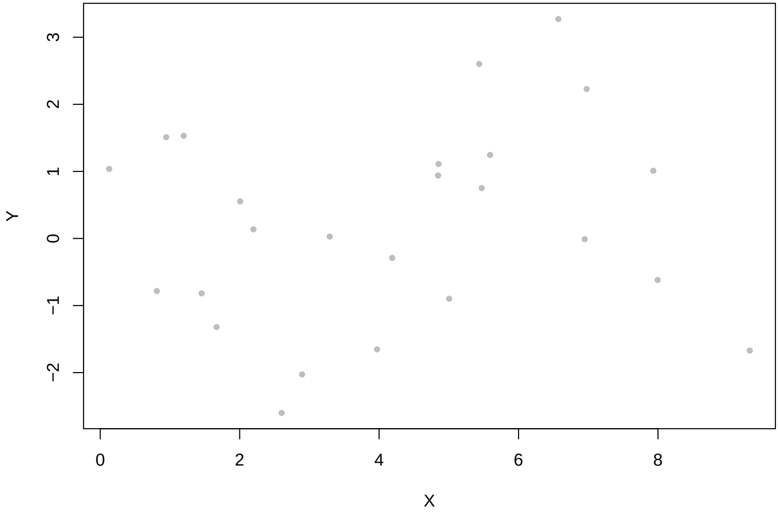
Readings

- Murphy (2012) *Machine Learning: a Probabilistic Perspective*
- James, Witten, Hastie and Tibshirani (2013) *An Introduction to Statistical Learning*

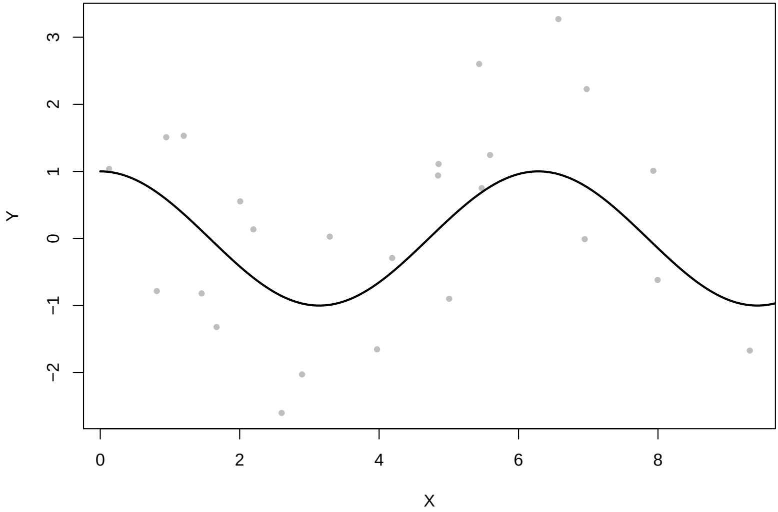
Readings

- Murphy (2012) *Machine Learning: a Probabilistic Perspective*
- James, Witten, Hastie and Tibshirani (2013) *An Introduction to Statistical Learning*
- Gelman and Hill (2008) *Data Analysis Using Regression and Multilevel/Hierarchical Models*

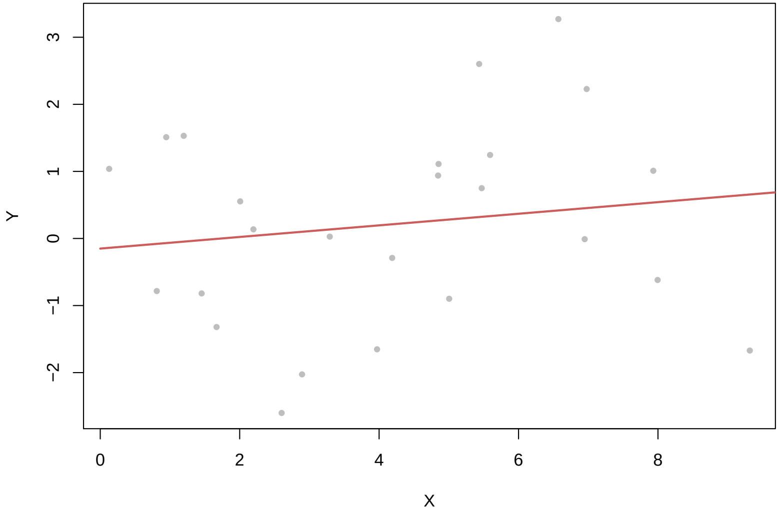
The Core Idea: Penalizing Complexity



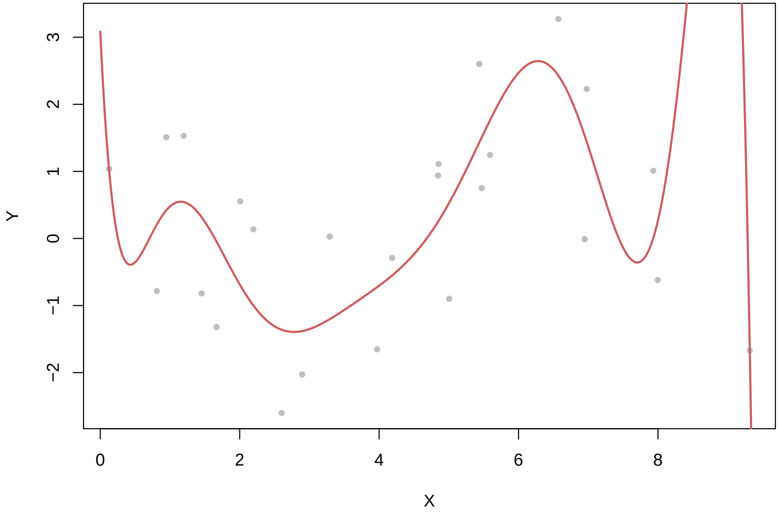
The Core Idea: Penalizing Complexity



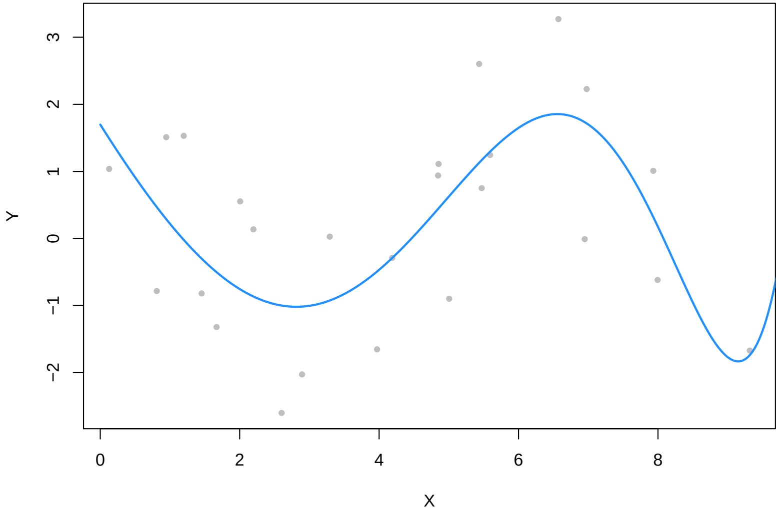
The Core Idea: Penalizing Complexity



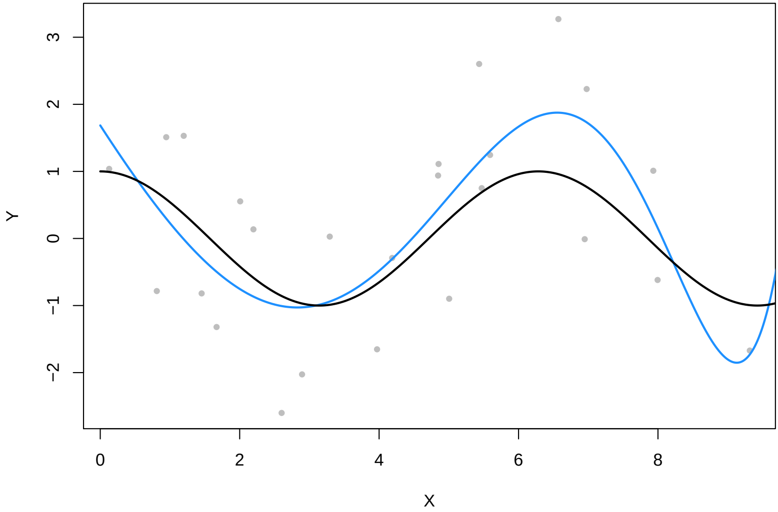
The Core Idea: Penalizing Complexity



The Core Idea: Penalizing Complexity



The Core Idea: Penalizing Complexity



Improving Estimation by Regularization

- Theme of this week is improving estimation through regularization or shrinkage

Improving Estimation by Regularization

- Theme of this week is improving estimation through regularization or shrinkage
- The core idea is that we may want to draw an estimate towards a particular point, generally inducing bias in exchange for a reduction in variance

Improving Estimation by Regularization

- Theme of this week is improving estimation through regularization or shrinkage
- The core idea is that we may want to draw an estimate towards a particular point, generally inducing bias in exchange for a reduction in variance
- Hierarchical models induce regularization to draw a set of group specific coefficients towards each other.

Improving Estimation by Regularization

- Theme of this week is improving estimation through regularization or shrinkage
- The core idea is that we may want to draw an estimate towards a particular point, generally inducing bias in exchange for a reduction in variance
- Hierarchical models induce regularization to draw a set of group specific coefficients towards each other.
- We will start with the simpler case of drawing coefficients towards zero (although later we will consider drawing estimates towards a data-driven point)

Prediction Accuracy and Interpretation

- Including regularization can improve the predictive accuracy of models, particularly in settings where n, the number of observations, is not much larger than p, the number of variables.

Prediction Accuracy and Interpretation

- Including regularization can improve the predictive accuracy of models, particularly in settings where n, the number of observations, is not much larger than p, the number of variables.
- When there are many variables with small or irrelevant effects, certain types of shrinkage can perform variable selection which zeroes out coefficients leaving only a small subset of variables.

Prediction Accuracy and Interpretation

- Including regularization can improve the predictive accuracy of models, particularly in settings where n, the number of observations, is not much larger than p, the number of variables.
- When there are many variables with small or irrelevant effects, certain types of shrinkage can perform variable selection which zeroes out coefficients leaving only a small subset of variables.
- Regularizers which draw coefficients to exact zeroes are called sparsity-inducing regularizers

Prediction Accuracy and Interpretation

- Including regularization can improve the predictive accuracy of models, particularly in settings where n, the number of observations, is not much larger than p, the number of variables.
- When there are many variables with small or irrelevant effects, certain types of shrinkage can perform variable selection which zeroes out coefficients leaving only a small subset of variables.
- Regularizers which draw coefficients to exact zeroes are called sparsity-inducing regularizers
- Regularization attempts to improve the generalizability of the model by penalizing extreme solutions even if they fit the current dataset better.

Mathematical Form of Regularization

Mathematical Form of Regularization

- In standard OLS we minimize the following criterion:

$$\text{RSS} = \sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right)^2$$

Mathematical Form of Regularization

- In standard OLS we minimize the following criterion:

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2$$

- In regularization we add a penalty such that we want to minimize:

$$\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

for some positive regularization penalty λ and a value q which determines the type of regularizer

Mathematical Form of Regularization

- In standard OLS we minimize the following criterion:

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2$$

- In regularization we add a penalty such that we want to minimize:

$$\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

  for some positive regularization penalty λ and a value q which determines the type of regularizer
- As the coefficients get larger, the penalty term increases

Mathematical Form of Regularization

- In standard OLS we minimize the following criterion:

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2$$

- In regularization we add a penalty such that we want to minimize:

$$\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \right) \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q$$

  for some positive regularization penalty λ and a value q which determines the type of regularizer

- As the coefficients get larger, the penalty term increases
- The math here is for least-squares but it also works for GLMs by replacing the RSS term with the negative log likelihood

Equivalent Views

- What is the penalty function doing?

Equivalent Views

- What is the penalty function doing?
- View 1: Penalizing Complex Functions

Equivalent Views

- What is the penalty function doing?
- View 1: Penalizing Complex Functions
  - instead of minimizing the loss, minimize the loss plus a complexity penalty

Equivalent Views

- What is the penalty function doing?
- View 1: Penalizing Complex Functions
    - instead of minimizing the loss, minimize the loss plus a complexity penalty
    - larger values of β imply a more complicated model (because a smaller change in X leads to a bigger change in y)

Equivalent Views

- What is the penalty function doing?
- View 1: Penalizing Complex Functions
  - instead of minimizing the loss, minimize the loss plus a complexity penalty
  - larger values of β imply a more complicated model (because a smaller change in X leads to a bigger change in y)
  - also allows estimation in settings where $p > n$

Equivalent Views

- What is the penalty function doing?
- View 1: Penalizing Complex Functions
  - instead of minimizing the loss, minimize the loss plus a complexity penalty
  - larger values of β imply a more complicated model (because a smaller change in X leads to a bigger change in y)
  - also allows estimation in settings where $p > n$
- View 2: Bayesian Prior

Equivalent Views

- What is the penalty function doing?
- View 1: Penalizing Complex Functions
  - instead of minimizing the loss, minimize the loss plus a complexity penalty
  - larger values of β imply a more complicated model (because a smaller change in X leads to a bigger change in y)
  - also allows estimation in settings where $p > n$
- View 2: Bayesian Prior
  - the prior distribution for β encodes values that we believe are a priori more reasonable

Equivalent Views

- What is the penalty function doing?
- View 1: Penalizing Complex Functions
  - instead of minimizing the loss, minimize the loss plus a complexity penalty
  - larger values of β imply a more complicated model (because a smaller change in X leads to a bigger change in y)
  - also allows estimation in settings where $p > n$
- View 2: Bayesian Prior
  - the prior distribution for β encodes values that we believe are a priori more reasonable
  - a prior distribution centered at 0 regularizes by penalizing larger values of β

Equivalent Views

- What is the penalty function doing?
- View 1: Penalizing Complex Functions
  - instead of minimizing the loss, minimize the loss plus a complexity penalty
  - larger values of β imply a more complicated model (because a smaller change in X leads to a bigger change in y)
  - also allows estimation in settings where $p > n$
- View 2: Bayesian Prior
  - the prior distribution for β encodes values that we believe are a priori more reasonable
  - a prior distribution centered at 0 regularizes by penalizing larger values of β
  - finding the maximum of the posterior (MAP inference) is equivalent to maximizing the likelihood with regularization

Bias-Variance Tradeoff

Bias-Variance Tradeoff

- At the heart of this is the bias-variance tradeoff

Bias-Variance Tradeoff

- At the heart of this is the bias-variance tradeoff
- Adding regularization can increase bias and in return reduce variance

Bias-Variance Tradeoff

- At the heart of this is the bias-variance tradeoff
- Adding regularization can increase bias and in return reduce variance
- We might care about minimizing the expected loss or expected prediction error

Bias-Variance Tradeoff

- At the heart of this is the bias-variance tradeoff
- Adding regularization can increase bias and in return reduce variance
- We might care about minimizing the expected loss or expected prediction error

$$\begin{aligned}\mathcal{R}(p(X,Y),f) &= E[L(Y,f(X))] \\ &= \int_{X\times Y} L(Y,f(X))p(X,Y)dXdY\end{aligned}$$

Bias-Variance Tradeoff

- At the heart of this is the bias-variance tradeoff
- Adding regularization can increase bias and in return reduce variance
- We might care about minimizing the expected loss or expected prediction error

$$\begin{aligned}\mathcal{R}(p(X,Y),f) &= E[L(Y,f(X))] \\ &= \int_{X\times Y} L(Y,f(X))p(X,Y)dXdY\end{aligned}$$

- How does bias and variance come into it?

Bias-Variance Tradeoff

- At the heart of this is the bias-variance tradeoff
- Adding regularization can increase bias and in return reduce variance
- We might care about minimizing the expected loss or expected prediction error

$$\begin{aligned}\mathcal{R}(p(X,Y),f) &= E[L(Y,f(X))] \\ &= \int_{X\times Y} L(Y,f(X))p(X,Y)dXdY\end{aligned}$$

- How does bias and variance come into it?
- Assume squared loss, and an estimated function $\hat{f}$, and fixed X's. The pointwise expected prediction error is:

$$\begin{aligned}\mathcal{R}(x_0) &= \mathbb{E}[(Y-\hat{f}(x_0))^2|X=x_0] \\ &= \sigma_\epsilon^2 + (\mathbb{E}[\hat{f}(x_0)] - f(x_0))^2 + \mathbb{E}[\hat{f}(x_0) - \mathbb{E}\hat{f}(x_0)]^2 \\ &= \text{Irreducible error} + \text{Bias}^2 + \text{Variance}\end{aligned}$$

Mean Square Error

Suppose θ is some value of the true parameter

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\mathrm{E}[(\hat{\theta} - \theta)^2]$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\text{E}[(\hat{\theta} - \theta)^2] = E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} \quad = \quad E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\begin{aligned} \mathsf{E}[(\hat{\theta} - \theta)^2] &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\ &= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \end{aligned}$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\begin{aligned}
\mathrm{E}[(\hat{\theta} - \theta)^2] &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + (E[\widehat{\theta} - \theta])^2
\end{aligned}$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\begin{aligned}
\text{E}[(\hat{\theta} - \theta)^2] &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + (E[\widehat{\theta} - \theta])^2 \\
&= \text{Var}(\theta) + \text{Bias}^2
\end{aligned}$$

Mean Square Error

Suppose θ is some value of the true parameter
Bias:

$$\text{Bias} = E[\widehat{\theta} - \theta]$$

We may care about average distance from truth

$$\begin{aligned}
\mathrm{E}[(\hat{\theta} - \theta)^2] &= E[\hat{\theta}^2] - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + E[\hat{\theta}]^2 - 2\theta E[\hat{\theta}] + \theta^2 \\
&= E[\hat{\theta}^2] - E[\hat{\theta}]^2 + (E[\widehat{\theta} - \theta])^2 \\
&= \text{Var}(\theta) + \text{Bias}^2
\end{aligned}$$

To reduce MSE, we are willing to induce bias to decrease variance $\rightsquigarrow$ methods that shrink coefficients toward zero

Two Canonical Regularizers

Two Canonical Regularizers

- There are an enormous number of regularizers for the squared error regression problem.

Two Canonical Regularizers

- There are an enormous number of regularizers for the squared error regression problem.
- They have really great names too:

Two Canonical Regularizers

- There are an enormous number of regularizers for the squared error regression problem.
- They have really great names too: tikhonov regularization, ridge regression,

Two Canonical Regularizers

- There are an enormous number of regularizers for the squared error regression problem.
- They have really great names too: tikhonov regularization, ridge regression, least absolute shrinkage and selection operator (LASSO),

Two Canonical Regularizers

- There are an enormous number of regularizers for the squared error regression problem.
- They have really great names too: tikhonov regularization, ridge regression, least absolute shrinkage and selection operator (LASSO), elastic net,

Two Canonical Regularizers

- There are an enormous number of regularizers for the squared error regression problem.
- They have really great names too: tikhonov regularization, ridge regression, least absolute shrinkage and selection operator (LASSO), elastic net, grouped lasso, fused lasso, adaptive lasso, gamma lasso, Bayesian lasso, square-root lasso, hierarchical adaptive lasso,

Two Canonical Regularizers

- There are an enormous number of regularizers for the squared error regression problem.
- They have really great names too: tikhonov regularization, ridge regression, least absolute shrinkage and selection operator (LASSO), elastic net, grouped lasso, fused lasso, adaptive lasso, gamma lasso, Bayesian lasso, square-root lasso, hierarchical adaptive lasso, smoothly clipped absolute deviation, horseshoe, bridge regression ...

Two Canonical Regularizers

- There are an enormous number of regularizers for the squared error regression problem.
- They have really great names too: tikhonov regularization, ridge regression, least absolute shrinkage and selection operator (LASSO), elastic net, grouped lasso, fused lasso, adaptive lasso, gamma lasso, Bayesian lasso, square-root lasso, hierarchical adaptive lasso, smoothly clipped absolute deviation, horseshoe, bridge regression ...
- We will cover two which come up frequently ridge regression and LASSO

Some Practical Matters and Notation

Some Practical Matters and Notation

- We will assume that covariates are standardized to have mean 0 and variance 1. This is to ensure that the different covariates are treated equivalently. We can always reproject them to their original scale.

Some Practical Matters and Notation

- We will assume that covariates are standardized to have mean 0 and variance 1. This is to ensure that the different covariates are treated equivalently. We can always reproject them to their original scale.
- We may talk about the penalty functions in terms of norms. The ℓ_2 norm is defined as $||\beta||_2 = \sqrt{\sum_{j=1}^{J} \beta_j^2}$

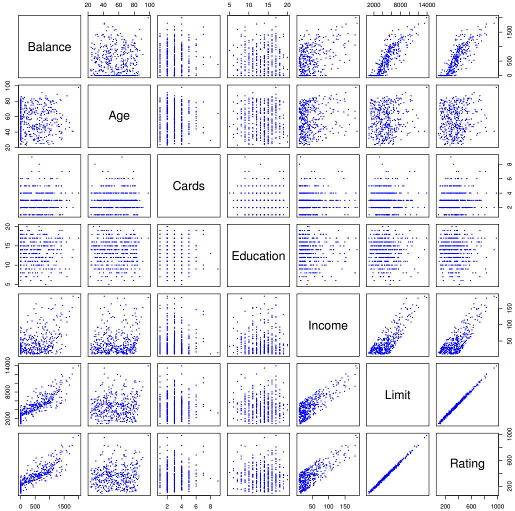
Some Practical Matters and Notation

- We will assume that covariates are standardized to have mean 0 and variance 1. This is to ensure that the different covariates are treated equivalently. We can always reproject them to their original scale.
- We may talk about the penalty functions in terms of norms. The ℓ_2 norm is defined as $||\beta||_2 = \sqrt{\sum_{j=1}^{J} \beta_j^2}$
- More generally we can define the ℓ_p norm as $||\beta||_p = \left(\sum_{j=1}^{J} |\beta_j|^p\right)^{\frac{1}{p}}$

Some Practical Matters and Notation

- We will assume that covariates are standardized to have mean 0 and variance 1. This is to ensure that the different covariates are treated equivalently. We can always reproject them to their original scale.
- We may talk about the penalty functions in terms of norms. The ℓ_2 norm is defined as $||\beta||_2 = \sqrt{\sum_{j=1}^{J} \beta_j^2}$
- More generally we can define the ℓ_p norm as $||\beta||_p = \left(\sum_{j=1}^{J} |\beta_j|^p\right)^{\frac{1}{p}}$
- We will use a running example of Credit Data which predicts credit card balance of a number of individuals using many predictors

Credit Data



Balance
Age
Cards
Education
Income
Limit
Rating
20 40 60 80 100
5 10 15 20
2000 8000 14000
0 500 1500
2 4 6 8
50 100 150
200 600 1000

Quadratic Regularizers

Penalty for model complexity

Quadratic Regularizers

Penalty for model complexity

$$f(\beta, \boldsymbol{X}, \boldsymbol{Y})$$

Quadratic Regularizers

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right) \right)^2$$

Quadratic Regularizers

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \;=\; \sum_{i=1}^{N}\left(y_i - \left(\beta_0 + \sum_{j=1}^{J}\beta_j x_{ij}\right)\right)^2 + \underbrace{\lambda\sum_{j=1}^{J}\beta_j^2}_{\text{Penalty}}$$

Quadratic Regularizers

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \;=\; \sum_{i=1}^{N}\left(y_i - \left(\beta_0 + \sum_{j=1}^{J}\beta_j x_{ij}\right)\right)^2 + \underbrace{\lambda\sum_{j=1}^{J}\beta_j^2}_{\text{Penalty}}$$

where:

Quadratic Regularizers

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N}\left(y_i - \left(\beta_0 + \sum_{j=1}^{J}\beta_j x_{ij}\right)\right)^2 + \lambda \underbrace{\sum_{j=1}^{J}\beta_j^2}_{\text{Penalty}}$$

where:

- $\beta_0 \rightsquigarrow$ intercept

Quadratic Regularizers

Penalty for model complexity

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) \;=\; \sum_{i=1}^{N} \left(y_i - \left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right) \right)^2 + \underbrace{\lambda \sum_{j=1}^{J} \beta_j^2}_{\text{Penalty}}$$

where:

- $\beta_0 \rightsquigarrow$ intercept
- $\lambda \rightsquigarrow$ penalty parameter

Ridge Regression $\rightsquigarrow$ Intuition (for a simple setting)

Suppose $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_J$.

Ridge Regression $\rightsquigarrow$ Intuition (for a simple setting)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_{J}$.

$$\widehat{\beta} \;=\; \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}$$

Ridge Regression $\rightsquigarrow$ Intuition (for a simple setting)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$\begin{aligned}\widehat{\beta} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\ &= \boldsymbol{X}^{'}\boldsymbol{Y}\end{aligned}$$

Ridge Regression $\rightsquigarrow$ Intuition (for a simple setting)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \boldsymbol{X}^{'}\boldsymbol{Y} \\
\boldsymbol{\beta}^{\text{ridge}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda \boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}
\end{aligned}$$

Ridge Regression $\rightsquigarrow$ Intuition (for a simple setting)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \boldsymbol{X}^{'}\boldsymbol{Y} \\
\boldsymbol{\beta}^{\text{ridge}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda\boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= (\boldsymbol{I}_j + \lambda\boldsymbol{I}_j)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}
\end{aligned}
$$

Ridge Regression $\rightsquigarrow$ Intuition (for a simple setting)

Suppose $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_J$.

$$
\begin{aligned}
\widehat{\beta} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} \\
&= \boldsymbol{X}'\boldsymbol{Y} \\
\beta^{\text{ridge}} &= \left(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I}_J\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} \\
&= (\boldsymbol{I}_j + \lambda\boldsymbol{I}_j)^{-1}\boldsymbol{X}'\boldsymbol{Y} \\
&= (\boldsymbol{I}_j + \lambda\boldsymbol{I}_j)^{-1}\widehat{\beta}
\end{aligned}
$$

Ridge Regression $\rightsquigarrow$ Intuition (for a simple setting)

Suppose $\boldsymbol{X}^{'}\boldsymbol{X} = \boldsymbol{I}_J$.

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \boldsymbol{X}^{'}\boldsymbol{Y} \\
\boldsymbol{\beta}^{\text{ridge}} &= \left(\boldsymbol{X}^{'}\boldsymbol{X} + \lambda\boldsymbol{I}_J\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \left(\boldsymbol{I}_j + \lambda\boldsymbol{I}_j\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\
&= \left(\boldsymbol{I}_j + \lambda\boldsymbol{I}_j\right)^{-1}\widehat{\boldsymbol{\beta}} \\
\beta_j^{\text{Ridge}} &= \frac{\widehat{\beta}_j}{1+\lambda}
\end{aligned}
$$

Selecting λ

$$\lambda \sum_j \beta_j^2$$

Selecting λ

$$\lambda \sum_j \beta_j^2$$

- λ controls the relative impact of the penalty term and the likelihood: selecting a good value is important!

Selecting λ

$$\lambda \sum_j \beta_j^2$$

- λ controls the relative impact of the penalty term and the likelihood: selecting a good value is important!
- Often referred to as a tuning parameter and most machine learning approaches have (at least) one

Selecting λ

$$\lambda \sum_j \beta_j^2$$

- λ controls the relative impact of the penalty term and the likelihood: selecting a good value is important!
- Often referred to as a tuning parameter and most machine learning approaches have (at least) one
- A higher value of λ indicates a lower tolerance for complexity the fitted model.

Selecting λ

$$\lambda \sum_j \beta_j^2$$

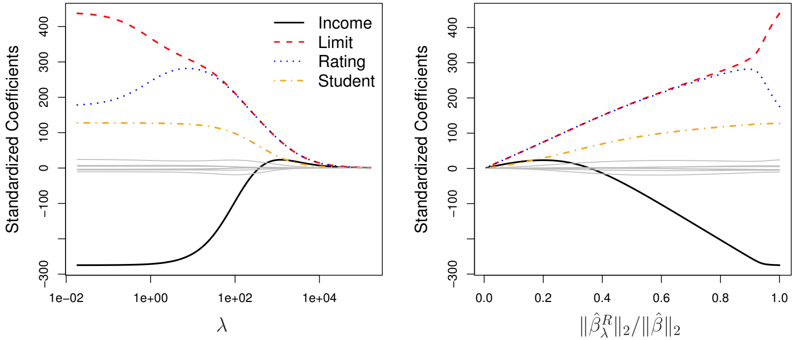
- λ controls the relative impact of the penalty term and the likelihood: selecting a good value is important!
- Often referred to as a tuning parameter and most machine learning approaches have (at least) one
- A higher value of λ indicates a lower tolerance for complexity the fitted model.
- We most often use cross-validation

Selecting λ

$$\lambda \sum_j \beta_j^2$$

- λ controls the relative impact of the penalty term and the likelihood: selecting a good value is important!
- Often referred to as a tuning parameter and most machine learning approaches have (at least) one
- A higher value of λ indicates a lower tolerance for complexity the fitted model.
- We most often use cross-validation
- We can visualize with a regularization path, a calculation across all values of λ

Regularization Path



Why Would We Use Ridge?

Why Would We Use Ridge?

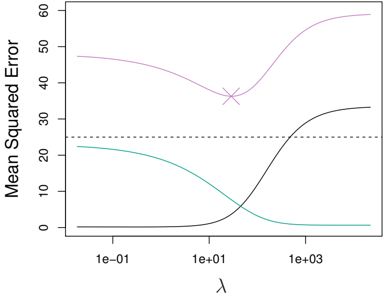
- As λ increases, the flexibility of the model decreases $\rightsquigarrow$ decreased variance, increased bias

Why Would We Use Ridge?

- As λ increases, the flexibility of the model decreases $\rightsquigarrow$ decreased variance, increased bias
- The trick is to find a place where the tradeoff is favorable

Why Would We Use Ridge?

- As λ increases, the flexibility of the model decreases $\rightsquigarrow$ decreased variance, increased bias
- The trick is to find a place where the tradeoff is favorable



Squared bias (black), variance (green), test mean squared error (purple). Dashed line is the minimum possible MSE

Lasso Regression Objective Function/Optimization

A different penalty $\rightsquigarrow$ different behavior

Lasso Regression Objective Function/Optimization

A different penalty $\rightsquigarrow$ different behavior

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right) \right)^2 + \lambda \sum_{j=1}^{J} \underbrace{|\beta_j|}_{\text{Penalty}}$$

Lasso Regression Objective Function/Optimization

A different penalty $\rightsquigarrow$ different behavior

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right) \right)^2 + \lambda \sum_{j=1}^{J} \underbrace{|\beta_j|}_{\text{Penalty}}$$

- Optimization is non-linear (due to the absolute value)

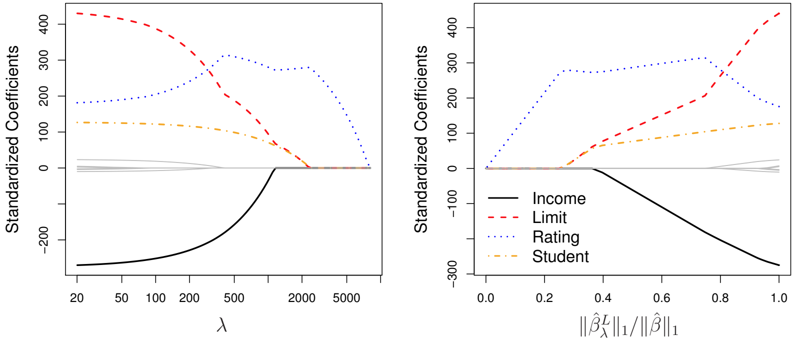
Lasso Regression Objective Function/Optimization

A different penalty $\rightsquigarrow$ different behavior

$$f(\boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{i=1}^{N} \left(y_i - \left(\beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} \right) \right)^2 + \lambda \sum_{j=1}^{J} \underbrace{|\beta_j|}_{\text{Penalty}}$$

- Optimization is non-linear (due to the absolute value)
- Induces sparsity $\rightsquigarrow$ sets some coefficients to zero

Regularization Path: Lasso



Lasso Regression $\rightsquigarrow$ Soft Thresholding

- In a simple special case where $X'X = I_J$, one can show that the LASSO update is:

Lasso Regression $\rightsquigarrow$ Soft Thresholding

- In a simple special case where $X'X = I_J$, one can show that the LASSO update is:

$$\beta_j^{\text{LASSO}} = \text{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_+$$

Lasso Regression $\rightsquigarrow$ Soft Thresholding

- In a simple special case where $X'X = I_J$, one can show that the LASSO update is:

$$\beta_j^{\text{LASSO}} \quad = \quad \text{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_+$$

  - where $\text{sign}(\cdot) \rightsquigarrow 1$ or -1
  - $\left(|\widehat{\beta}_j| - \lambda\right)_+ = \max(|\widehat{\beta}_j| - \lambda, 0)$

Lasso Regression $\rightsquigarrow$ Soft Thresholding

- In a simple special case where $X'X = I_J$, one can show that the LASSO update is:

$$\beta_j^{\text{LASSO}} = \text{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_+$$

  - where $\text{sign}(\cdot) \rightsquigarrow 1$ or -1
  - $\left(|\widehat{\beta}_j| - \lambda\right)_+ = \max(|\widehat{\beta}_j| - \lambda, 0)$
- Thus up to a particular value the coefficient remains 0.

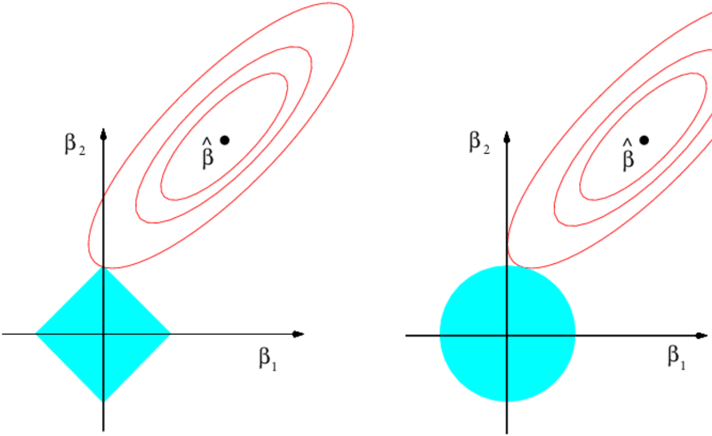
Lasso Regression $\leadsto$ Soft Thresholding

- In a simple special case where $X'X = I_J$, one can show that the LASSO update is:

$$\beta_j^{\text{LASSO}} = \text{sign}\left(\widehat{\beta}_j\right)\left(|\widehat{\beta}_j| - \lambda\right)_+$$

  - where $\text{sign}(\cdot) \leadsto 1$ or -1
  - $\left(|\widehat{\beta}_j| - \lambda\right)_+ = \max(|\widehat{\beta}_j| - \lambda, 0)$
- Thus up to a particular value the coefficient remains 0.
- Where does the sparsity come from? and why doesn't ridge have it?

Lasso vs. Ridge



Origins of Sparsity

- It turns out the sparsity is deeply connected to the fact that the penalty term is not differentiable. This also makes optimization difficult

Origins of Sparsity

- It turns out the sparsity is deeply connected to the fact that the penalty term is not differentiable. This also makes optimization difficult
- One intuition is that the marginal rate of penalization is constant as you move away from zero, but grows under the ridge penalty.

Origins of Sparsity

- It turns out the sparsity is deeply connected to the fact that the penalty term is not differentiable. This also makes optimization difficult
- One intuition is that the marginal rate of penalization is constant as you move away from zero, but grows under the ridge penalty.
- In the special cases we saw different types of shrinkage: ridge shrinks each estimate by the same proportion, Lasso shrinks each estimate by the same amount

Origins of Sparsity

- It turns out the sparsity is deeply connected to the fact that the penalty term is not differentiable. This also makes optimization difficult
- One intuition is that the marginal rate of penalization is constant as you move away from zero, but grows under the ridge penalty.
- In the special cases we saw different types of shrinkage: ridge shrinks each estimate by the same proportion, Lasso shrinks each estimate by the same amount
- Let's do a quick mathematical example

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

$$\sum_{j=1}^{2} \beta_j^2 \quad = \quad \frac{1}{2} + \frac{1}{2} = 1$$

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

$$\begin{aligned} \sum_{j=1}^{2} \beta_j^2 &= \frac{1}{2} + \frac{1}{2} = 1 \\ \sum_{j=1}^{2} \tilde{\beta}_j^2 &= 1 + 0 = 1 \end{aligned}$$

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

$$\begin{aligned}\sum_{j=1}^{2} \beta_j^2 &= \frac{1}{2} + \frac{1}{2} = 1 \\ \sum_{j=1}^{2} \tilde{\beta}_j^2 &= 1 + 0 = 1\end{aligned}$$

Under LASSO

Comparing Ridge and LASSO

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

$$\begin{aligned}
\sum_{j=1}^{2} \beta_j^2 &= \frac{1}{2} + \frac{1}{2} = 1 \\
\sum_{j=1}^{2} \tilde{\beta}_j^2 &= 1 + 0 = 1
\end{aligned}$$

Under LASSO

$$\sum_{j=1}^{2} |\beta_j| = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2}$$

Comparing Ridge and LASSO

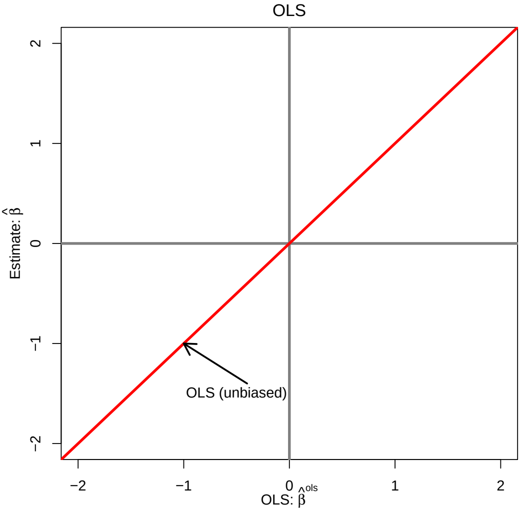
Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Under ridge:

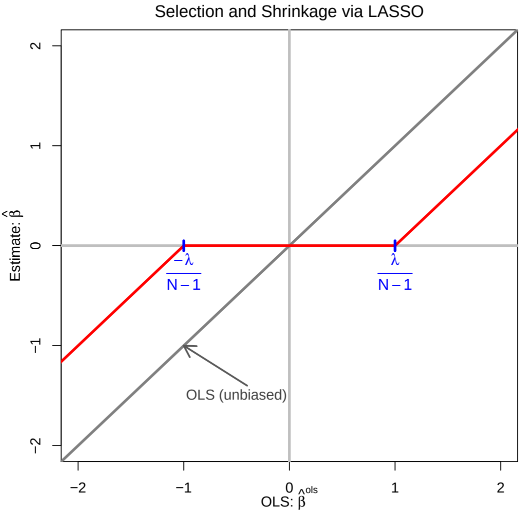
$$\begin{aligned} \sum_{j=1}^{2} \beta_j^2 &= \frac{1}{2} + \frac{1}{2} = 1 \\ \sum_{j=1}^{2} \tilde{\beta}_j^2 &= 1 + 0 = 1 \end{aligned}$$

Under LASSO

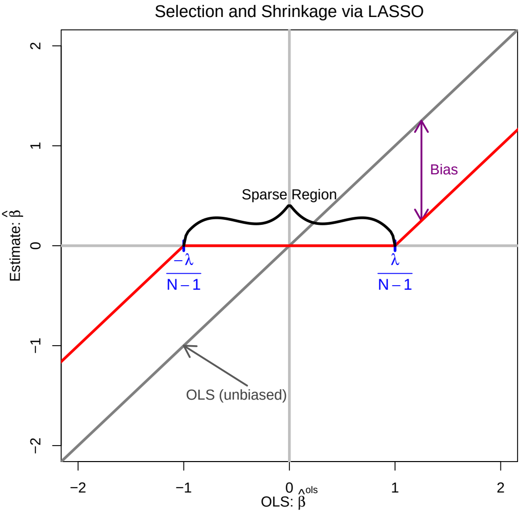
$$\begin{aligned} \sum_{j=1}^{2} |\beta_j| &= \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2} \\ \sum_{j=1}^{2} |\tilde{\beta}_j| &= 1 + 0 = 1 \end{aligned}$$



$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2$$



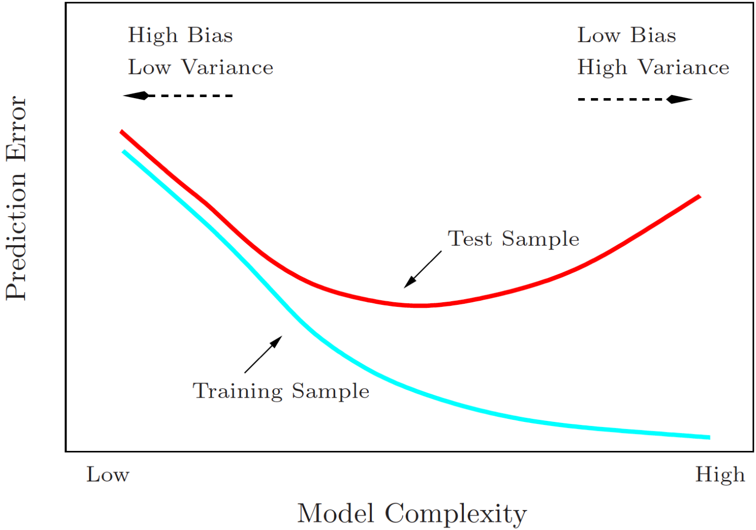
$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda|\beta|$$



$$\widehat{\beta} = \arg\min_\beta \sum_{i=1}^{N} (Y_i - X_i\beta)^2 + \lambda|\beta|$$

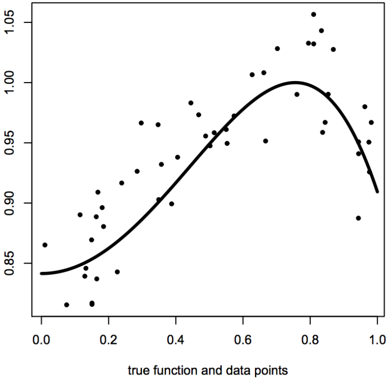
Bias-Variance Tradeoff in Action

Bias-Variance Tradeoff in Action



Example Synthetic Problem[2]

$$y = \sin(1 + x^2) + \epsilon$$



[2] These slides are adapted from material by Radford Neal.

Linear Basis Function Models

Linear Basis Function Models

- We talked before about polynomials x^2, x^3, x^4 for modeling non-linearities, this is a linear basis function model.

Linear Basis Function Models

- We talked before about polynomials x^2, x^3, x^4 for modeling non-linearities, this is a linear basis function model.
- In general the idea is to do a linear regression of y on $\phi_1(x), \phi_2(x), \ldots, \phi_{m-1}(x)$ where ϕ_j are basis functions.

Linear Basis Function Models

- We talked before about polynomials x^2, x^3, x^4 for modeling non-linearities, this is a linear basis function model.
- In general the idea is to do a linear regression of y on $\phi_1(x), \phi_2(x), \ldots, \phi_{m-1}(x)$ where ϕ_j are basis functions.
- The model is now:

$$y = f(x, \beta) + \epsilon$$

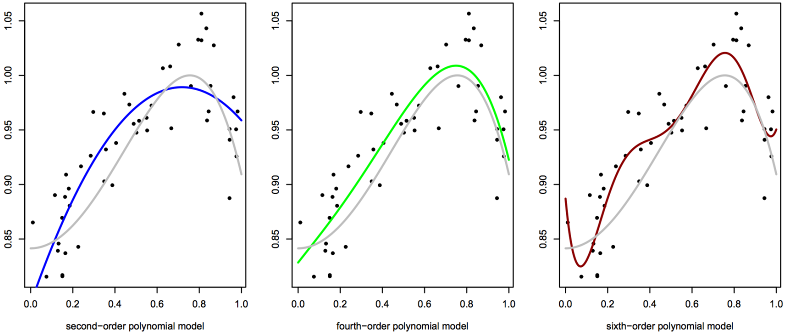
$$f(x, \beta) = \beta_0 + \sum_{j=1}^{m-1} \beta_j \phi_j(x) = \beta^T \phi(x)$$

Polynomial Basis Functions

We can look at OLS fits with polynomial basis functions of increasing order.

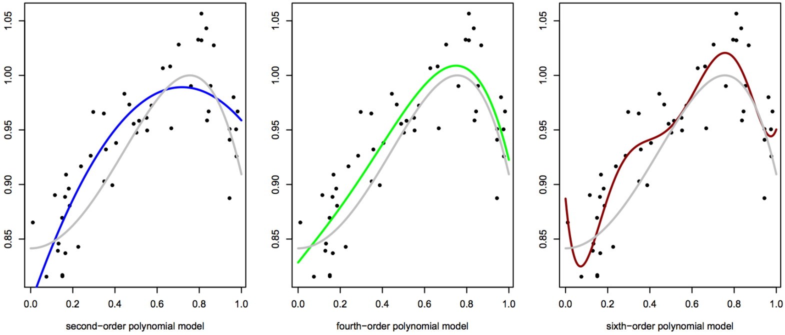
Polynomial Basis Functions

We can look at OLS fits with polynomial basis functions of increasing order.



Polynomial Basis Functions

We can look at OLS fits with polynomial basis functions of increasing order.



It appears that the last model is too complex and is overfitting a bit.

Local Basis Functions

Local Basis Functions

Polynomials are global basis functions, each affecting the prediction over the whole input space. Often local basis functions are more appropriate.

Local Basis Functions

Polynomials are global basis functions, each affecting the prediction over the whole input space. Often local basis functions are more appropriate.

One choice is a Gaussian basis function

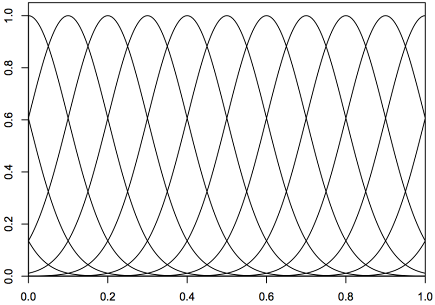
$$\phi_j(x) = \exp(-(x-\mu_j)^2)/2s^2)$$

Local Basis Functions

Polynomials are global basis functions, each affecting the prediction over the whole input space. Often local basis functions are more appropriate.

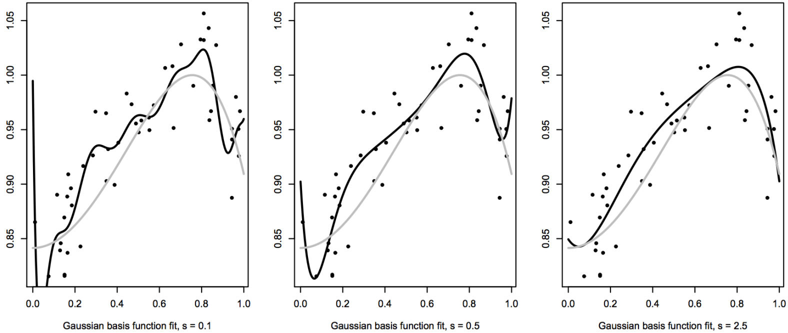
One choice is a Gaussian basis function

$$\phi_j(x) = \exp(-(x-\mu_j)^2)/2s^2)$$



Gaussian basis functions, s = 0.1

Gaussian Basis Fits



Regularization

Regularization

- We've seen that flexible models can lead to overfitting

Regularization

- We've seen that flexible models can lead to overfitting
- Two ways to address: limit model flexibility or use a flexible model and regularize

Regularization

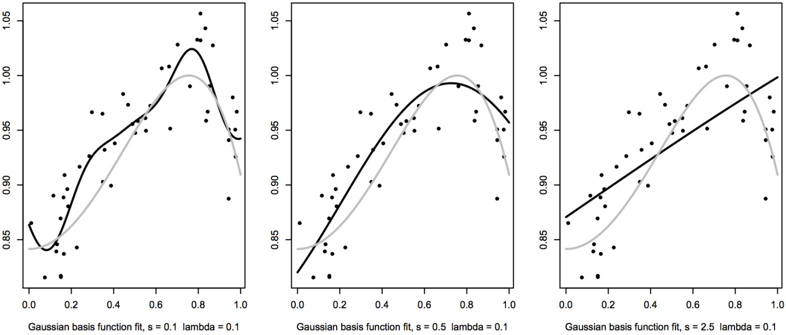
- We've seen that flexible models can lead to overfitting
- Two ways to address: limit model flexibility or use a flexible model and regularize
- Regularization is the way to express preference for smoothness in our function

Regularization

- We've seen that flexible models can lead to overfitting
- Two ways to address: limit model flexibility or use a flexible model and regularize
- Regularization is the way to express preference for smoothness in our function
- Let's look at the ridge penalty $\lambda \sum_{j=1}^{m-1} \beta_j^2$ where λ controls the strength of the penalty.

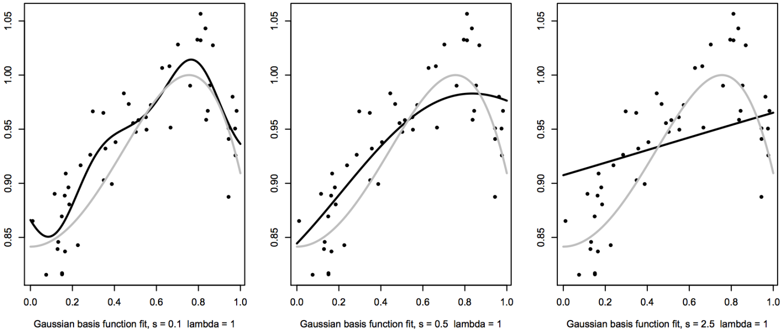
Results

Here are the results with $\lambda = 0.1$:



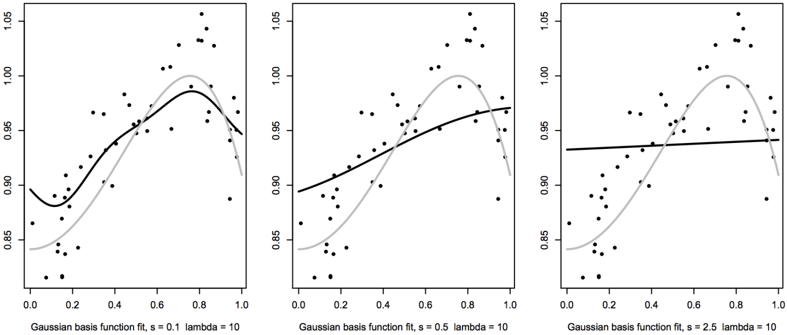
Results

Here are the results with $\lambda = 1$:



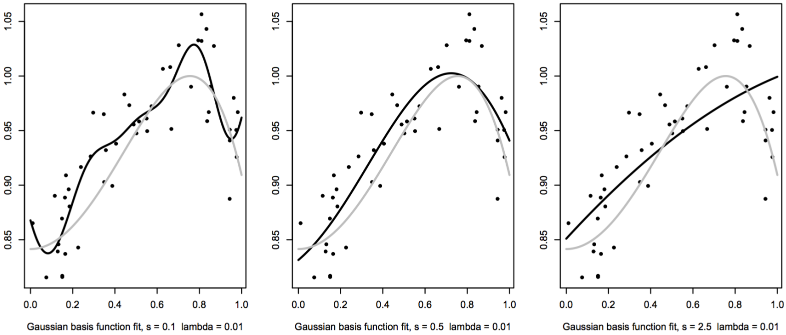
Results

Here are the results with $\lambda = 10$:



Results

Here are the results with $\lambda = 0.01$:



We've Seen Ridge Regression Before

- Generalized Additive Models (GAM's) from the `mgcv` package use ridge regression.

We've Seen Ridge Regression Before

- Generalized Additive Models (GAM's) from the `mgcv` package use ridge regression.
- Also recall Kernel Regularized Least Squares (KRLS) Hainmueller and Hazlett (2013). "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach" *Political Analysis*.

Paper

Ratkovic and Tingley (2017) "Sparse Estimation and Uncertainty with Application to Subgroup Analysis" *Political Analysis*.

Motivating Example: Subgroup Analysis

Moving past average treatment effects

Motivating Example: Subgroup Analysis

Moving past average treatment effects

- effects for different groups of individuals (subgroups)
- the effect of combinations of treatments

Motivating Example: Subgroup Analysis

Moving past average treatment effects

- effects for different groups of individuals (subgroups)
- the effect of combinations of treatments

Increasingly complex designs

- conjoint analysis
- repeated observations

Motivating Example: Subgroup Analysis

Moving past average treatment effects

- effects for different groups of individuals (subgroups)
- the effect of combinations of treatments

Increasingly complex designs

- conjoint analysis
- repeated observations

Proliferation of possible effects

Multiple hypothesis testing

Multiple hypothesis testing

Majors problems with multiple hypothesis testing

Multiple hypothesis testing

Majors problems with multiple hypothesis testing

- p-values become uninformative!
- $1,000$ possible subsets: 50 false positives!
- Publication bias

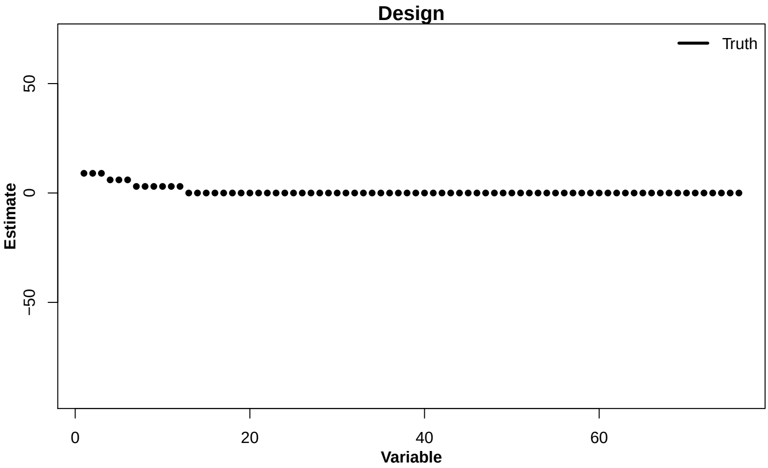
Multiple hypothesis testing

Majors problems with multiple hypothesis testing

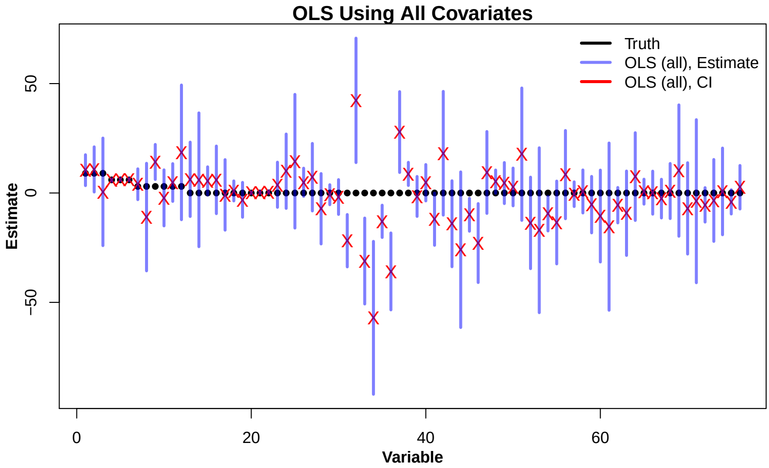
- p-values become uninformative!
- $1,000$ possible subsets: 50 false positives!
- Publication bias

Subsetting data still requires specification hunting, and is also underpowered.

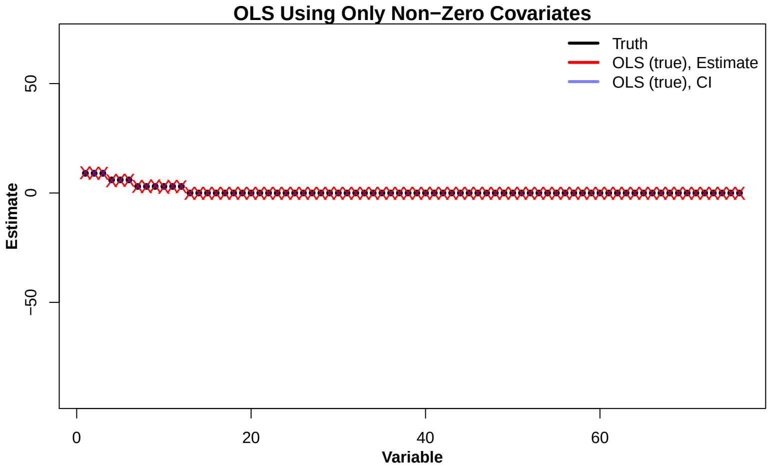
Design: $N = 10{,}000$; $K = 76$



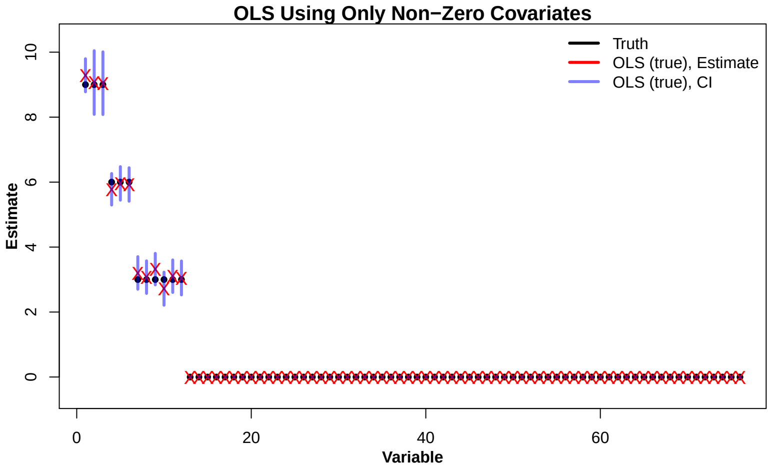
OLS


OLS Using All Covariates
Truth
OLS (all), Estimate
OLS (all), CI
Estimate
50
0
−50
0
20
40
60
Variable

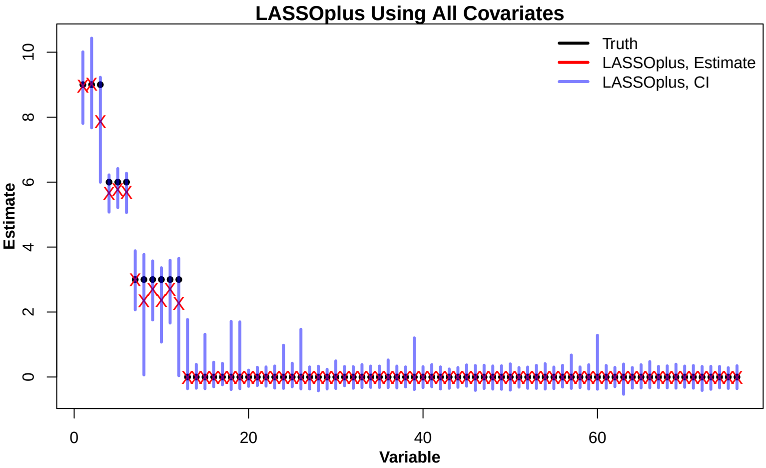
"Oracle:" OLS on only non-zero effects



"Oracle:" Zooming in



LASSOplus



LASSOplus Using All Covariates
Truth
LASSOplus, Estimate
LASSOplus, CI
Estimate
Variable
0
2
4
6
8
10
0
20
40
60

Introducing LASSOplus

Introducing LASSOplus

Statistical properties

- Sparse estimates

Introducing LASSOplus

Statistical properties

- Sparse estimates
- Oracle property

Introducing LASSOplus

Statistical properties

- Sparse estimates
- Oracle property
  - Consistent variable selection
  - Asymptotically equivalent to model fit to non-zero effects

Introducing LASSOplus

Statistical properties

- Sparse estimates
- Oracle property
  - Consistent variable selection
  - Asymptotically equivalent to model fit to non-zero effects
- Frequentist coverage

Introducing LASSOplus

Statistical properties

- Sparse estimates
- Oracle property
  - Consistent variable selection
  - Asymptotically equivalent to model fit to non-zero effects
- Frequentist coverage

Practical properties

- Easy to implement

```
s1<-sparsereg(y, X, cbind(t1, t2), scale.type="TX", EM=TRUE)
```

Introducing LASSOplus

Statistical properties

- Sparse estimates
- Oracle property
  - Consistent variable selection
  - Asymptotically equivalent to model fit to non-zero effects
- Frequentist coverage

Practical properties

- Easy to implement

```
s1<-sparsereg(y, X, cbind(t1, t2), scale.type="TX", EM=TRUE)
```

- Flexibility
  - Up to three-way random effects
  - Continuous and binary outcomes

Contributions of LASSOplus

Statistical contributions

1. Weakly informative prior structure
2. Sparse estimates
3. Approximate confidence intervals
4. Oracle property

Contributions of LASSOplus

Statistical contributions

1. Weakly informative prior structure
2. Sparse estimates
3. Approximate confidence intervals
4. Oracle property

Practical contributions

1. Pre-processes data
2. Handles repeated observations
3. Extends beyond standard linear model (probit, tobit, etc.)

Contributions of LASSOplus

Statistical contributions

1. Weakly informative prior structure
2. Sparse estimates
3. Approximate confidence intervals
4. Oracle property

Practical contributions

1. Pre-processes data
2. Handles repeated observations
3. Extends beyond standard linear model (probit, tobit, etc.)

All implemented in `sparsereg` (Ratkovic and Tingley 2015) in **R**.

A Simple Example

A hypothetical experiment

A Simple Example

A hypothetical experiment

- Two treatments
  - $T_1 \in \{a, b, c\}$
  - $T_2 \in \{a, b, c, d\}$

A Simple Example

A hypothetical experiment

- Two treatments
  - $T_1 \in \{a, b, c\}$
  - $T_2 \in \{a, b, c, d\}$
- Pre-treatment covariates: $[X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}]$

A Simple Example

A hypothetical experiment

- Two treatments
  - $T_1 \in \{a, b, c\}$
  - $T_2 \in \{a, b, c, d\}$
- Pre-treatment covariates: $[X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}]$
- Data generating process

$$
\begin{aligned}
Y_i = & 3 + 2 \cdot X_{i2} + 2 \cdot \mathbf{1}(T_{i1} = a) - 2 \cdot \mathbf{1}(T_{i1} = b) \\
& - 2 \cdot X_{i2} \cdot \mathbf{1}(T_{i1} = b) + 2 \cdot X_{i2} \cdot \mathbf{1}(T_{i2} = c) + \epsilon_i
\end{aligned}
$$

where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 4)$; $N = 500$

A Simple Example

A hypothetical experiment

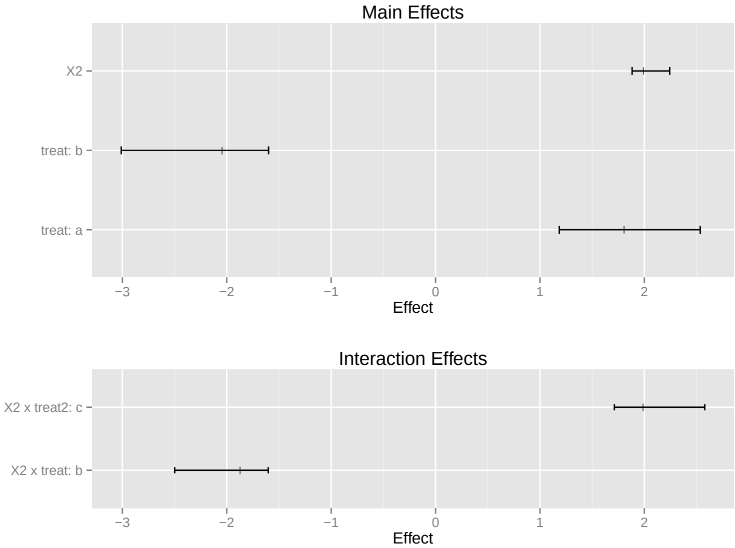
- Two treatments
  - $T_1 \in \{a, b, c\}$
  - $T_2 \in \{a, b, c, d\}$
- Pre-treatment covariates: $[X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}]$
- Data generating process

$$\begin{aligned} Y_i = & 3 + 2 \cdot X_{i2} + 2 \cdot \mathbf{1}(T_{i1} = a) - 2 \cdot \mathbf{1}(T_{i1} = b) \\ & - 2 \cdot X_{i2} \cdot \mathbf{1}(T_{i1} = b) + 2 \cdot X_{i2} \cdot \mathbf{1}(T_{i2} = c) + \epsilon_i \end{aligned}$$

  where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 4)$; $N = 500$

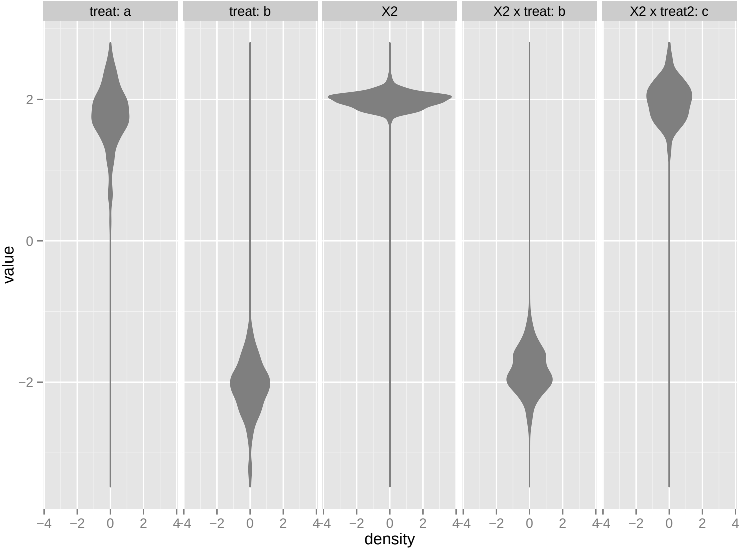
- $K = \underbrace{3 + 4 + 5}_{\text{main effects}} + \underbrace{5 \cdot (3 + 4)}_{\text{interaction terms}} = 47$

plot(s1)



Main Effects
X2
treat: b
treat: a
−3
−2
−1
0
1
2
Effect
Interaction Effects
X2 x treat2: c
X2 x treat: b
−3
−2
−1
0
1
2
Effect

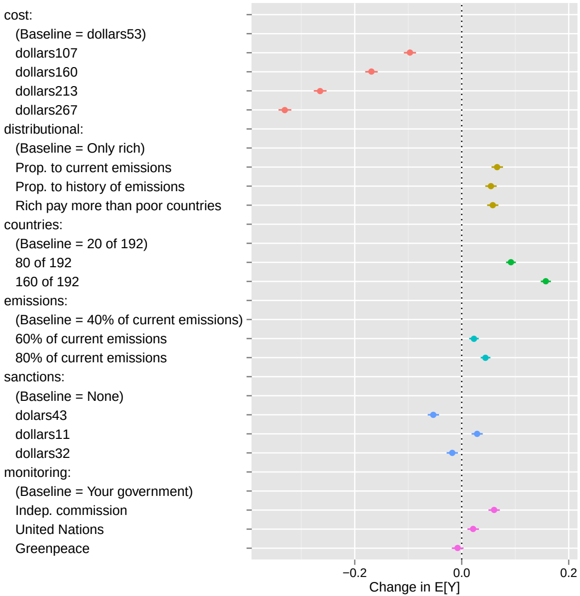
violinplot(s1)

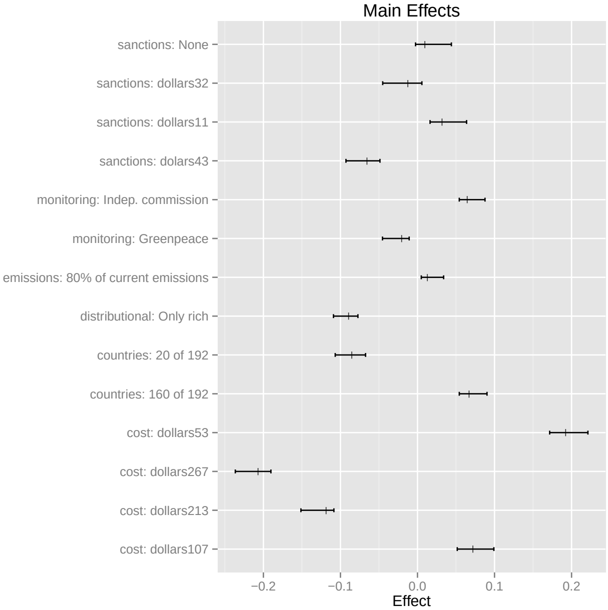


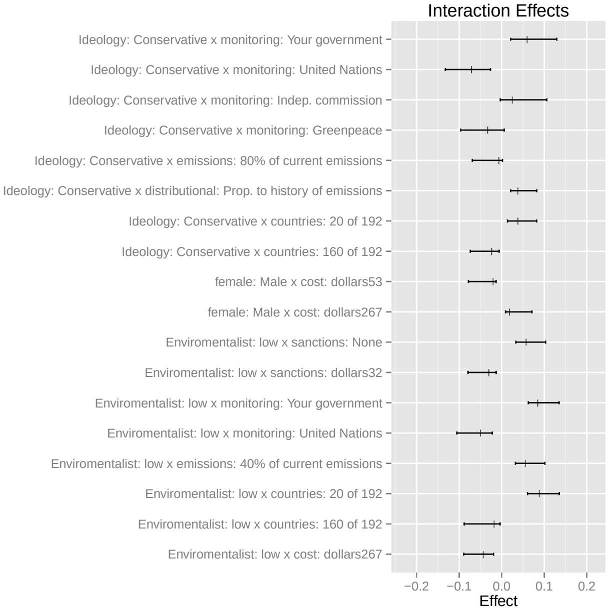
Application

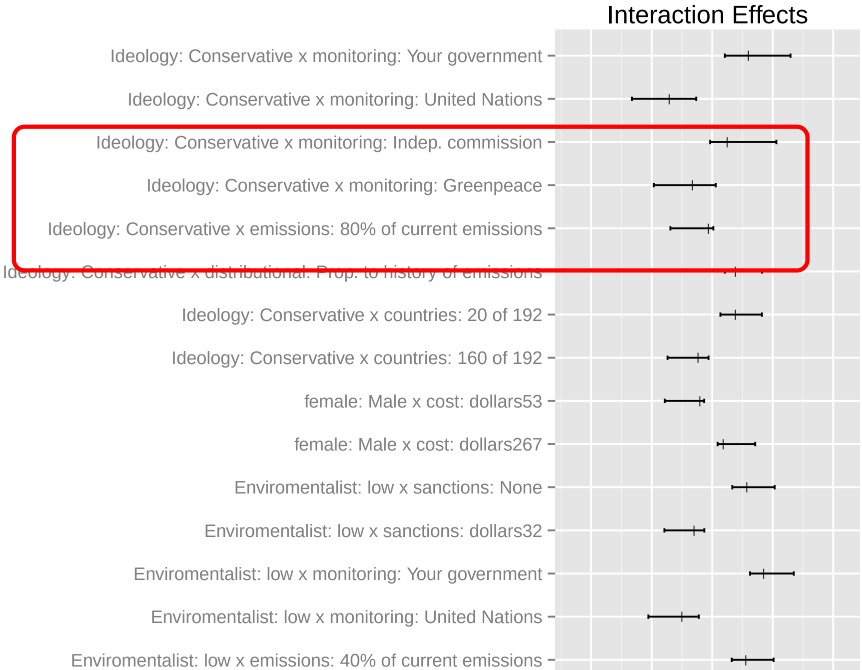
Study: Bechtel and Scheve 2013

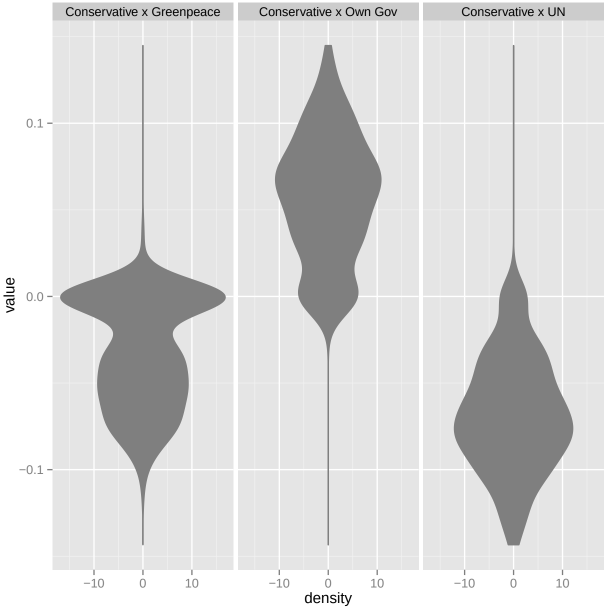
- Effect of international treaty on climate design on support
- Conjoint experiment across four countries
- Treatment conditions: cost, extent of other countries participating, extent of sanctions, who monitors
- 215 total effects
  - 16 covariates; 31 main effects; 6 treatments, 23 levels; 184 treatment $\times$ covariate effects
- Investigated sub-group effects by multiple split sample analyses


cost:
(Baseline = dollars53)
dollars107
dollars160
dollars213
dollars267
distributional:
(Baseline = Only rich)
Prop. to current emissions
Prop. to history of emissions
Rich pay more than poor countries
countries:
(Baseline = 20 of 192)
80 of 192
160 of 192
emissions:
(Baseline = 40% of current emissions)
60% of current emissions
80% of current emissions
sanctions:
(Baseline = None)
dolars43
dollars11
dollars32
monitoring:
(Baseline = Your government)
Indep. commission
United Nations
Greenpeace
−0.2
0.0
0.2
Change in E[Y]


Main Effects
sanctions: None
sanctions: dollars32
sanctions: dollars11
sanctions: dolars43
monitoring: Indep. commission
monitoring: Greenpeace
emissions: 80% of current emissions
distributional: Only rich
countries: 20 of 192
countries: 160 of 192
cost: dollars53
cost: dollars267
cost: dollars213
cost: dollars107
−0.2
−0.1
0.0
0.1
0.2
Effect


Interaction Effects
Ideology: Conservative x monitoring: Your government
Ideology: Conservative x monitoring: United Nations
Ideology: Conservative x monitoring: Indep. commission
Ideology: Conservative x monitoring: Greenpeace
Ideology: Conservative x emissions: 80% of current emissions
Ideology: Conservative x distributional: Prop. to history of emissions
Ideology: Conservative x countries: 20 of 192
Ideology: Conservative x countries: 160 of 192
female: Male x cost: dollars53
female: Male x cost: dollars267
Enviromentalist: low x sanctions: None
Enviromentalist: low x sanctions: dollars32
Enviromentalist: low x monitoring: Your government
Enviromentalist: low x monitoring: United Nations
Enviromentalist: low x emissions: 40% of current emissions
Enviromentalist: low x countries: 20 of 192
Enviromentalist: low x countries: 160 of 192
Enviromentalist: low x cost: dollars267
−0.2
−0.1
0.0
0.1
0.2
Effect


Interaction Effects
Ideology: Conservative x monitoring: Your government
Ideology: Conservative x monitoring: United Nations
Ideology: Conservative x monitoring: Indep. commission
Ideology: Conservative x monitoring: Greenpeace
Ideology: Conservative x emissions: 80% of current emissions
Ideology: Conservative x distributional: Prop. to history of emissions
Ideology: Conservative x countries: 20 of 192
Ideology: Conservative x countries: 160 of 192
female: Male x cost: dollars53
female: Male x cost: dollars267
Enviromentalist: low x sanctions: None
Enviromentalist: low x sanctions: dollars32
Enviromentalist: low x monitoring: Your government
Enviromentalist: low x monitoring: United Nations
Enviromentalist: low x emissions: 40% of current emissions


Conservative x Greenpeace
Conservative x Own Gov
Conservative x UN
value
0.1
0.0
−0.1
−10
0
10
density

Conclusion on LASSOplus

LASSOplus is an estimator that

1. possesses the Oracle property
2. achieves a low FDR
3. identifies non-zero effects
4. returns approximate confidence intervals

Where we are

Where we are

- Regularization/shrinkage as pulling coefficients towards 0

Where we are

- Regularization/shrinkage as pulling coefficients towards 0
- Our goal was to reduce variance at the possible expense of bias

Where we are

- Regularization/shrinkage as pulling coefficients towards 0
- Our goal was to reduce variance at the possible expense of bias
- In general hierarchical models we use reguarlization in order to share information across related units

Where we are

- Regularization/shrinkage as pulling coefficients towards 0
- Our goal was to reduce variance at the possible expense of bias
- In general hierarchical models we use reguarlization in order to share information across related units
- Let's consider a single example of a hierarchical model: eight schools

Eight Schools Background

- ETS analyzes special coaching program on test scores

Eight Schools Background

- ETS analyzes special coaching program on test scores
- 8 separate parallel experiments in different high schools

Eight Schools Background

- ETS analyzes special coaching program on test scores
- 8 separate parallel experiments in different high schools
- Outcome was the score on a special administration of SAT-V with scores varying between 200 and 800 ($\mu = 500, \sigma = 100$)

Eight Schools Background

- ETS analyzes special coaching program on test scores
- 8 separate parallel experiments in different high schools
- Outcome was the score on a special administration of SAT-V with scores varying between 200 and 800 ($\mu = 500, \sigma = 100$)
- SAT is designed to be resistant to short-term efforts intended to boost performance, but each school thought it was a success.

Eight Schools Background

- ETS analyzes special coaching program on test scores
- 8 separate parallel experiments in different high schools
- Outcome was the score on a special administration of SAT-V with scores varying between 200 and 800 ($\mu = 500, \sigma = 100$)
- SAT is designed to be resistant to short-term efforts intended to boost performance, but each school thought it was a success.
- No prior reason to believe that one program would be more effective than the others

Eight Schools Background

- ETS analyzes special coaching program on test scores
- 8 separate parallel experiments in different high schools
- Outcome was the score on a special administration of SAT-V with scores varying between 200 and 800 ($\mu = 500, \sigma = 100$)
- SAT is designed to be resistant to short-term efforts intended to boost performance, but each school thought it was a success.
- No prior reason to believe that one program would be more effective than the others
- Treatment effects estimated controlling for PSAT-M and PSAT-V scores

Eight Schools Background

- ETS analyzes special coaching program on test scores
- 8 separate parallel experiments in different high schools
- Outcome was the score on a special administration of SAT-V with scores varying between 200 and 800 ($\mu = 500, \sigma = 100$)
- SAT is designed to be resistant to short-term efforts intended to boost performance, but each school thought it was a success.
- No prior reason to believe that one program would be more effective than the others
- Treatment effects estimated controlling for PSAT-M and PSAT-V scores
- A bit over the 30 students in each school

Eight Schools Background

- ETS analyzes special coaching program on test scores
- 8 separate parallel experiments in different high schools
- Outcome was the score on a special administration of SAT-V with scores varying between 200 and 800 ($\mu = 500, \sigma = 100$)
- SAT is designed to be resistant to short-term efforts intended to boost performance, but each school thought it was a success.
- No prior reason to believe that one program would be more effective than the others
- Treatment effects estimated controlling for PSAT-M and PSAT-V scores
- A bit over the 30 students in each school
- For the sake of scale: an 8-point increase in the score indicates about 1 more test item was answered correctly.

Eight Schools Background

- ETS analyzes special coaching program on test scores
- 8 separate parallel experiments in different high schools
- Outcome was the score on a special administration of SAT-V with scores varying between 200 and 800 ($\mu = 500, \sigma = 100$)
- SAT is designed to be resistant to short-term efforts intended to boost performance, but each school thought it was a success.
- No prior reason to believe that one program would be more effective than the others
- Treatment effects estimated controlling for PSAT-M and PSAT-V scores
- A bit over the 30 students in each school
- For the sake of scale: an 8-point increase in the score indicates about 1 more test item was answered correctly.
- (Analysis is from Rubin 1981, treatment via Gelman et al 2015)

Eight Schools Data

School	Est. Effect	SE
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Eight Schools Data

School	Est. Effect	SE
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Policy Question: What is the effect size in School A?

What do we know?

- Unbiased estimate: 28 points

What do we know?

- Unbiased estimate: 28 points
- Hypothesis test fails to reject hypothesis that true effect is the same for all of them

What do we know?

- Unbiased estimate: 28 points
- Hypothesis test fails to reject hypothesis that true effect is the same for all of them
- Should we analyze them all together? All separately?

What do we know?

- Unbiased estimate: 28 points
- Hypothesis test fails to reject hypothesis that true effect is the same for all of them
- Should we analyze them all together? All separately?
- It is the "same course" in every school, but they are different schools.

Options for Analysis

There are two clear options:

1. an unpooled analysis in which we use separate estimates for every school- in this case directly from the table

Options for Analysis

There are two clear options:

1. an unpooled analysis in which we use separate estimates for every school- in this case directly from the table
    - 2 moderate effects, 4 small effects and 2 small negative effects

Options for Analysis

There are two clear options:

1. an unpooled analysis in which we use separate estimates for every school- in this case directly from the table
   - 2 moderate effects, 4 small effects and 2 small negative effects
   - standard errors are large, 95% intervals overlap substantially

Options for Analysis

There are two clear options:

1. an unpooled analysis in which we use separate estimates for every school- in this case directly from the table
   - 2 moderate effects, 4 small effects and 2 small negative effects
   - standard errors are large, 95% intervals overlap substantially
2. a pooled analysis that generates a single estimate for all schools

Options for Analysis

There are two clear options:

1. an unpooled analysis in which we use separate estimates for every school- in this case directly from the table
   - 2 moderate effects, 4 small effects and 2 small negative effects
   - standard errors are large, 95% intervals overlap substantially
2. a pooled analysis that generates a single estimate for all schools
   - assume that all effects are exactly the same

Options for Analysis

There are two clear options:

1. an unpooled analysis in which we use separate estimates for every school- in this case directly from the table
   - 2 moderate effects, 4 small effects and 2 small negative effects
   - standard errors are large, 95% intervals overlap substantially
2. a pooled analysis that generates a single estimate for all schools
   - assume that all effects are exactly the same
   - we get the single effect size and standard error with inverse variance weighting of the unpooled estimates.

$$\bar{y}_{\cdot} = \frac{\sum_{j=1}^{8} \frac{1}{\sigma_j^2} \bar{y}_j}{\sum_{j=1}^{8} \frac{1}{\sigma_j^2}}$$

$$\sigma_{\cdot}^2 = \left(\sum_{j=1}^{8} \frac{1}{\sigma_j^2} \right)^{-1}$$

Options for Analysis

There are two clear options:

1. an unpooled analysis in which we use separate estimates for every school- in this case directly from the table
   - 2 moderate effects, 4 small effects and 2 small negative effects
   - standard errors are large, 95% intervals overlap substantially
2. a pooled analysis that generates a single estimate for all schools
   - assume that all effects are exactly the same
   - we get the single effect size and standard error with inverse variance weighting of the unpooled estimates.

$$\bar{y}_{.} = \frac{\sum_{j=1}^{8} \frac{1}{\sigma_j^2} \bar{y}_j}{\sum_{j=1}^{8} \frac{1}{\sigma_j^2}}$$

$$\sigma_{.}^2 = \left(\sum_{j=1}^{8} \frac{1}{\sigma_j^2} \right)^{-1}$$

   - the pooled estimate is 7.7 with standard error of 4.1. Thus the confidence interval is $[-.5, 15.9]$

Problems with Separate and Pooled Analysis

- The two approaches radically different results for school A: 28.4 (s.e. 14.9) vs. 7.7 (s.e. 4.1)

Problems with Separate and Pooled Analysis

- The two approaches radically different results for school A: 28.4 (s.e. 14.9) vs. 7.7 (s.e. 4.1)
- Under a Bayesian framework, the separate analysis implies the probability statement "the probability is $\frac{1}{2}$ that the true effect in A is more than 28.4"

Problems with Separate and Pooled Analysis

- The two approaches radically different results for school A: 28.4 (s.e. 14.9) vs. 7.7 (s.e. 4.1)
- Under a Bayesian framework, the separate analysis implies the probability statement "the probability is $\frac{1}{2}$ that the true effect in A is more than 28.4"
- This seems . . . dubious given the other results (remember we had no reason to believe one school would perform stronger than the others)

Problems with Separate and Pooled Analysis

- The two approaches radically different results for school A: 28.4 (s.e. 14.9) vs. 7.7 (s.e. 4.1)
- Under a Bayesian framework, the separate analysis implies the probability statement "the probability is $\frac{1}{2}$ that the true effect in A is more than 28.4"
- This seems ... dubious given the other results (remember we had no reason to believe one school would perform stronger than the others)
- The pooled analysis implies the statement "the probability is $\frac{1}{2}$ that the true effect in A is less than 7.7", it also implies that "the probability is $\frac{1}{2}$ that the true effect in A is less than the true effect in C"

Problems with Separate and Pooled Analysis

- The two approaches radically different results for school A: 28.4 (s.e. 14.9) vs. 7.7 (s.e. 4.1)
- Under a Bayesian framework, the separate analysis implies the probability statement "the probability is $\frac{1}{2}$ that the true effect in A is more than 28.4"
- This seems ... dubious given the other results (remember we had no reason to believe one school would perform stronger than the others)
- The pooled analysis implies the statement "the probability is $\frac{1}{2}$ that the true effect in A is less than 7.7", it also implies that "the probability is $\frac{1}{2}$ that the true effect in A is less than the true effect in C"
- Again these seem unlikely given the data

Borrowing Information

- We want an estimate that combines information from the 8 experiments without assuming that all the effects are equal

Borrowing Information

- We want an estimate that combines information from the 8 experiments without assuming that all the effects are equal
- Rubin suggests a middle path: a hierarchical model in which we

Borrowing Information

- We want an estimate that combines information from the 8 experiments without assuming that all the effects are equal
- Rubin suggests a middle path: a hierarchical model in which we
    1. assume that each school's true effect is drawn a Normal distribution with some unknown mean and standard deviation

Borrowing Information

- We want an estimate that combines information from the 8 experiments without assuming that all the effects are equal
- Rubin suggests a middle path: a hierarchical model in which we
    1. assume that each school's true effect is drawn a Normal distribution with some unknown mean and standard deviation
    2. assume that the observed effect in each school is sampled from a normal distribution with a mean equal to its true effect and standard deviation given in the table

Borrowing Information

- We want an estimate that combines information from the 8 experiments without assuming that all the effects are equal
- Rubin suggests a middle path: a hierarchical model in which we
  1. assume that each school's true effect is drawn a Normal distribution with some unknown mean and standard deviation
  2. assume that the observed effect in each school is sampled from a normal distribution with a mean equal to its true effect and standard deviation given in the table
- This model contains both the separate and pooled estimates as limiting special cases. If we force the standard deviation of the true effects to be zero, then all school get the same estimate, if we let it go to infinity we get the separate estimates

The Model

$$\bar{y}_j|\theta_j \sim \text{Normal}(\theta_j, \sigma_j^2)$$
$$\theta_j|\mu, \tau \sim \text{Normal}(\mu, \tau^2)$$
$$p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau)$$

Known: $\bar{y}_j, \sigma_j^2$
Unknown: τ, μ, θ

A General Hierarchical Model Form

First Stage: p(data | process, parameters)
Second Stage: p(process | parameters)
Third Stage: hyperparameters

A General Hierarchical Model Form

First Stage: p(data | process, parameters)
Second Stage: p(process | parameters)
Third Stage: hyperparameters

$$
\begin{aligned}
Y|X,\beta &\sim N(X\beta, \Sigma_Y) \\
\beta|Z,\alpha &\sim N(z\alpha, \Sigma_\beta) \\
\alpha &\sim N(\alpha_0, \Sigma_\alpha)
\end{aligned}
$$

Some Mechanics

How do the calculations work conditional on some values of the hyperparameters?

Some Mechanics

How do the calculations work conditional on some values of the hyperparameters?

The θs are latent variables which have a distribution. In Bayesian statistics we call this the posterior distribution.

Some Mechanics

How do the calculations work conditional on some values of the hyperparameters?

The θs are latent variables which have a distribution. In Bayesian statistics we call this the posterior distribution.

$$\theta_j|\mu, \tau, y \sim \mathrm{N}(\hat{\theta}_j, V_j)$$

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\bar{y}_j + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

$$V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

What is Happening?

- We are borrowing information between the schools

What is Happening?

- We are borrowing information between the schools
- Alternatively- we are regularizing estimates of the individual effects towards their grand mean

What is Happening?

- We are borrowing information between the schools
- Alternatively- we are regularizing estimates of the individual effects towards their grand mean
- This captures our intuition that while School A might have a larger effect, it is perhaps an overestimate

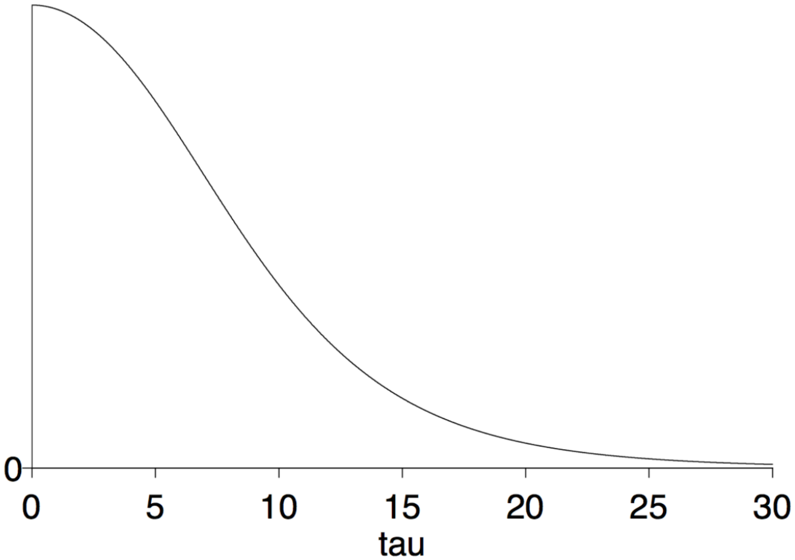
What is Happening?

- We are borrowing information between the schools
- Alternatively- we are regularizing estimates of the individual effects towards their grand mean
- This captures our intuition that while School A might have a larger effect, it is perhaps an overestimate
- The form show us that the amount of shrinkage is relative to our certainty about the estimate and how much we believe the individual effects matter

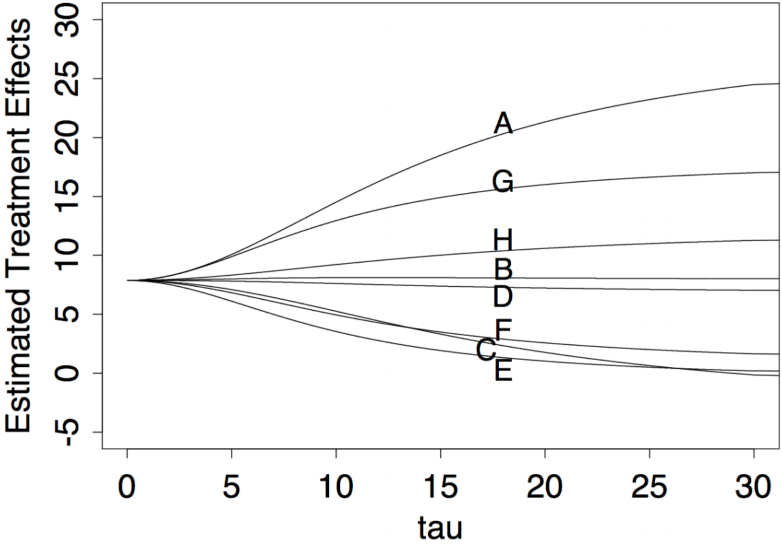
What is Happening?

- We are borrowing information between the schools
- Alternatively- we are regularizing estimates of the individual effects towards their grand mean
- This captures our intuition that while School A might have a larger effect, it is perhaps an overestimate
- The form show us that the amount of shrinkage is relative to our certainty about the estimate and how much we believe the individual effects matter
- Our final guess is that the median effect for school A is about 10 points with 50% probability between 7 and 16

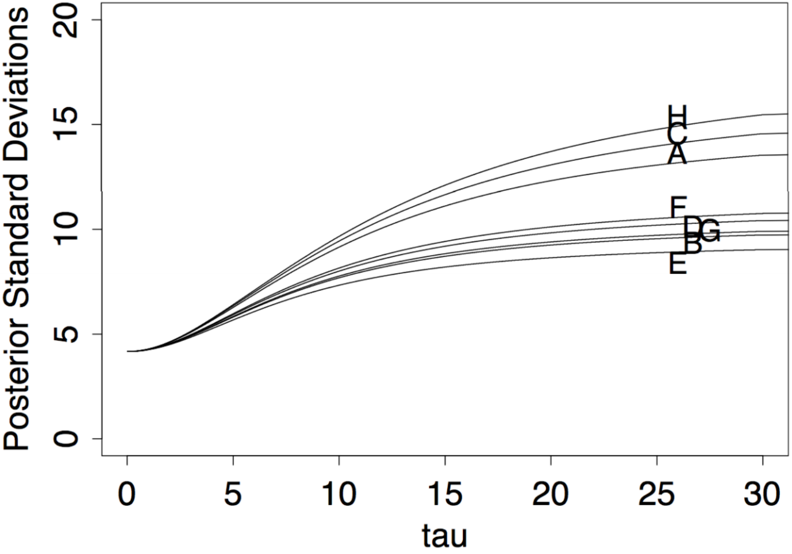
Results



Results

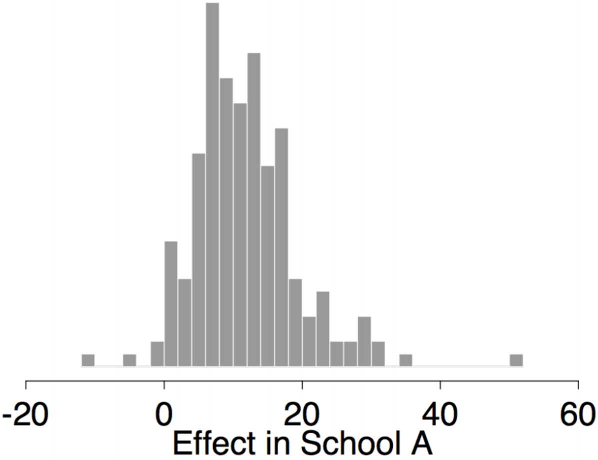

Estimated Treatment Effects
-5
0
5
10
15
20
25
30
A
G
H
B
D
F
C
E
0
5
10
15
20
25
30
tau

Results



Posterior Standard Deviations
0
5
10
15
20
tau
0
5
10
15
20
25
30
H
C
A
F
D
G
B
E

Results



The Great Thing About Eight Schools

- This is a microcosm of hierarchical modeling

The Great Thing About Eight Schools

- This is a microcosm of hierarchical modeling
- Works well when we have a decent number of groups and the individual group sample sizes are lowish

The Great Thing About Eight Schools

- This is a microcosm of hierarchical modeling
- Works well when we have a decent number of groups and the individual group sample sizes are lowish
- Allows us to capture variability in our treatment effects, variances etc.

The Great Thing About Eight Schools

- This is a microcosm of hierarchical modeling
- Works well when we have a decent number of groups and the individual group sample sizes are lowish
- Allows us to capture variability in our treatment effects, variances etc.
- Allows us to model dependence in our error terms

Beyond Eight Schools

- Eight Schools is a simple example without any covariates (sort of) and with the individual data abstracted away

Beyond Eight Schools

- Eight Schools is a simple example without any covariates (sort of) and with the individual data abstracted away
- Today we will consider the broader class of multilevel models

Beyond Eight Schools

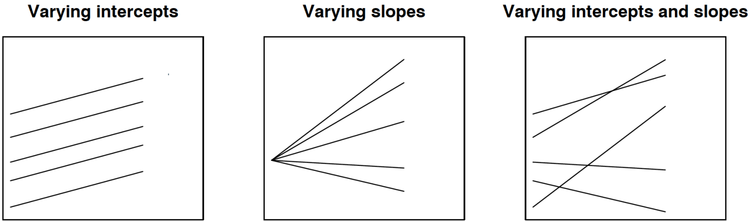
- Eight Schools is a simple example without any covariates (sort of) and with the individual data abstracted away
- Today we will consider the broader class of multilevel models
- Let's start with a simple structure: individuals within a group, individual level predictors only.

Beyond Eight Schools

- Eight Schools is a simple example without any covariates (sort of) and with the individual data abstracted away
- Today we will consider the broader class of multilevel models
- Let's start with a simple structure: individuals within a group, individual level predictors only.
- We can think of three model variants:

$$\begin{aligned}
\text{varying-intercept model:} y_i &= a_{j[i]} + \beta x_i + \epsilon_i \\
\text{varying-slope model:} y_i &= \alpha + \beta_{j[i]} x_i + \epsilon_i \\
\text{varying intercept and slope model:} y_i &= a_{j[i]} + \beta_{j[i]} x_i + \epsilon_i
\end{aligned}$$

Varying Intercept and Slopes



Example Data

ID	dad age	mom race	informal support	city ID	city name	enforce intensity	benefit level	city indicators 1	2	⋯	20
1	19	hisp	1	1	Oakland	0.52	1.01	1	0	⋯	0
2	27	black	0	1	Oakland	0.52	1.01	1	0	⋯	0
3	26	black	1	1	Oakland	0.52	1.01	1	0	⋯	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
248	19	white	1	3	Baltimore	0.05	1.10	0	0	⋯	0
249	26	black	1	3	Baltimore	0.05	1.10	0	0	⋯	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
1366	21	black	1	20	Norfolk	−0.11	1.08	0	0	⋯	1
1367	28	hisp	0	20	Norfolk	−0.11	1.08	0	0	⋯	1

Four ways to analyze

1. Individual-level regression:

Four ways to analyze

1. Individual-level regression:
   - include all the individual and city-level variables in the table

Four ways to analyze

1. Individual-level regression:
    - include all the individual and city-level variables in the table
    - restriction: can't capture city-level variation beyond the city level predictors

Four ways to analyze

1. Individual-level regression:
   - include all the individual and city-level variables in the table
   - restriction: can't capture city-level variation beyond the city level predictors
2. Group-level regression on city averages

Four ways to analyze

1. Individual-level regression:
   - include all the individual and city-level variables in the table
   - restriction: can't capture city-level variation beyond the city level predictors
2. Group-level regression on city averages
   - explain the average level outcome based on group-level covariates and individual-level average

Four ways to analyze

1. Individual-level regression:
    - include all the individual and city-level variables in the table
    - restriction: can't capture city-level variation beyond the city level predictors
2. Group-level regression on city averages
    - explain the average level outcome based on group-level covariates and individual-level average
    - restriction: fewer data points and removes ability of individual predictors to predict individual outcomes.

Four ways to analyze

1. Individual-level regression:
    - include all the individual and city-level variables in the table
    - restriction: can't capture city-level variation beyond the city level predictors
2. Group-level regression on city averages
    - explain the average level outcome based on group-level covariates and individual-level average
    - restriction: fewer data points and removes ability of individual predictors to predict individual outcomes.
3. Two-step analysis

Four ways to analyze

1. Individual-level regression:
   - include all the individual and city-level variables in the table
   - restriction: can't capture city-level variation beyond the city level predictors
2. Group-level regression on city averages
   - explain the average level outcome based on group-level covariates and individual-level average
   - restriction: fewer data points and removes ability of individual predictors to predict individual outcomes.
3. Two-step analysis
   - fit an logistic regression with individual variables and city level intercepts, in a second linear regression fit the estimated intercepts with group level covariates

Four ways to analyze

1. Individual-level regression:
   - include all the individual and city-level variables in the table
   - restriction: can't capture city-level variation beyond the city level predictors
2. Group-level regression on city averages
   - explain the average level outcome based on group-level covariates and individual-level average
   - restriction: fewer data points and removes ability of individual predictors to predict individual outcomes.
3. Two-step analysis
   - fit an logistic regression with individual variables and city level intercepts, in a second linear regression fit the estimated intercepts with group level covariates
   - retstriction: problems with small sample sizes, ignores individual/group variable interactions, ignores estimation uncertainty

Four ways to analyze

1. Individual-level regression:
   - include all the individual and city-level variables in the table
   - restriction: can't capture city-level variation beyond the city level predictors
2. Group-level regression on city averages
   - explain the average level outcome based on group-level covariates and individual-level average
   - restriction: fewer data points and removes ability of individual predictors to predict individual outcomes.
3. Two-step analysis
   - fit an logistic regression with individual variables and city level intercepts, in a second linear regression fit the estimated intercepts with group level covariates
   - retstriction: problems with small sample sizes, ignores individual/group variable interactions, ignores estimation uncertainty
4. Multilevel models

An Example Multilevel Model for Fragile Families

An Example Multilevel Model for Fragile Families

$$\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta + \alpha_{j[i]})$$
$$\alpha_j \sim \mathcal{N}(U_j\gamma, \sigma^2_\alpha)$$

where X are individual predictors, U are group predictors, and σ_a is the standard deviation of unexplained city-level variation.

An Example Multilevel Model for Fragile Families

$$\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta + \alpha_{j[i]})$$
$$\alpha_j \sim \mathcal{N}(U_j\gamma, \sigma^2_\alpha)$$

where X are individual predictors, U are group predictors, and σ_a is the standard deviation of unexplained city-level variation.

What does it mean to include group level predictors U?

An Example Multilevel Model for Fragile Families

$$\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta + \alpha_{j[i]})$$
$$\alpha_j \sim \mathcal{N}(U_j\gamma, \sigma^2_\alpha)$$

where X are individual predictors, U are group predictors, and σ_a is the standard deviation of unexplained city-level variation.

What does it mean to include group level predictors U?

In eight schools we saw partial pooling to the grand mean, here we see partial pooling to the regression prediction.

An Example Multilevel Model for Fragile Families

$$\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta + \alpha_{j[i]})$$
$$\alpha_j \sim \mathcal{N}(U_j\gamma, \sigma^2_\alpha)$$

where X are individual predictors, U are group predictors, and σ_a is the standard deviation of unexplained city-level variation.

What does it mean to include group level predictors U?

In eight schools we saw partial pooling to the grand mean, here we see partial pooling to the regression prediction.

The multilevel estimate of α_j is a weighted average of the no-pooling estimate for the group and the regression prediction.

Non-nested Structures

We can extend this framework to settings which are not cleanly nested such as longitudinal data.

Non-nested Structures

We can extend this framework to settings which are not cleanly nested such as longitudinal data.

person		parents smoke?		wave 1		wave 2		⋯
ID	sex	mom	dad	age	smokes?	age	smokes?	
1	f	Y	Y	15:0	N	15:6	N	⋯
2	f	N	N	14:7	N	15:1	N	⋯
3	m	Y	N	15:1	N	15:7	Y	⋯
4	f	N	N	15:3	N	15:9	N	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Non-nested Structures

We can extend this framework to settings which are not cleanly nested such as longitudinal data.

person ID	sex	parents smoke? mom	dad	wave 1 age	smokes?	wave 2 age	smokes?	...
1	f	Y	Y	15:0	N	15:6	N	...
2	f	N	N	14:7	N	15:1	N	...
3	m	Y	N	15:1	N	15:7	Y	...
4	f	N	N	15:3	N	15:9	N	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

$$\Pr(y_i = 1) = \text{logit}^{-1}(\beta_0 + \beta_1 \text{psmoke}_j + \beta_2 \text{female}_j + \beta_3 t + \beta_4(\text{female}_j)t + \alpha_j)$$

Fixed and Random Effects Nomenclature

Fixed and Random Effects Nomenclature

- Five wildly different uses

Fixed and Random Effects Nomenclature

- Five wildly different uses
- Rules of thumb:

Fixed and Random Effects Nomenclature

- Five wildly different uses
- Rules of thumb:
    - "fixed effects regression" means a regression with group level intercepts included as dummy variables (no shrinkage)
    - "fixed effects" within the former type are the group level intercepts
    - "fixed effects" within a multilevel/hierarchical model are the terms which don't vary by group

Fixed and Random Effects Nomenclature

- Five wildly different uses
- Rules of thumb:
  - "fixed effects regression" means a regression with group level intercepts included as dummy variables (no shrinkage)
  - "fixed effects" within the former type are the group level intercepts
  - "fixed effects" within a multilevel/hierarchical model are the terms which don't vary by group
- Perspectives and estimation in econometrics

Reasons to Do Multilevel Modeling

- Accounting for individual/group variation in estimating group-level coefficients

Reasons to Do Multilevel Modeling

- Accounting for individual/group variation in estimating group-level coefficients
- Modeling variation in individual-level regression coefficients

Reasons to Do Multilevel Modeling

- Accounting for individual/group variation in estimating group-level coefficients
- Modeling variation in individual-level regression coefficients
- Partially pooling to estimate regression coefficients for individual groups

Reasons to Do Multilevel Modeling

- Accounting for individual/group variation in estimating group-level coefficients
- Modeling variation in individual-level regression coefficients
- Partially pooling to estimate regression coefficients for individual groups

When should we bother?

Reasons to Do Multilevel Modeling

- Accounting for individual/group variation in estimating group-level coefficients
- Modeling variation in individual-level regression coefficients
- Partially pooling to estimate regression coefficients for individual groups

When should we bother?
Roughly speaking when the number of groups is > 5 with decent amounts of variation between groups and/or small group sizes.

Different Ways of Writing the Same Model

Different Ways of Writing the Same Model

- $\alpha_j \sim N(\mu_\alpha, \sigma^2_\alpha)$ can be rewritten as $\alpha_j = \mu_\alpha + \eta_j$ where $\eta_j \sim N(0, \sigma^2_\alpha)$

Different Ways of Writing the Same Model

- $\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$ can be rewritten as $\alpha_j = \mu_\alpha + \eta_j$ where $\eta_j \sim N(0, \sigma_\alpha^2)$
- This formulation leads naturally into expressing the model as a regression with multiple error terms:
  $y_i = X_i\beta + \mu_\alpha + \eta_j + \sigma_\epsilon^2$

Different Ways of Writing the Same Model

- $\alpha_j \sim N(\mu_\alpha, \sigma^2_\alpha)$ can be rewritten as $\alpha_j = \mu_\alpha + \eta_j$ where $\eta_j \sim N(0, \sigma^2_\alpha)$
- This formulation leads naturally into expressing the model as a regression with multiple error terms:
  $y_i = X_i\beta + \mu_\alpha + \eta_j + \sigma^2_\epsilon$
- We can also express it as a standard regression with correlated errors: $y_i = X_i\beta + \epsilon_i^{\text{all}}, \epsilon_i^{\text{all}} \sim N(0, \Sigma)$ where Σ is structured in a particular way.

Different Ways of Writing the Same Model

- $\alpha_j \sim N(\mu_\alpha, \sigma^2_\alpha)$ can be rewritten as $\alpha_j = \mu_\alpha + \eta_j$ where $\eta_j \sim N(0, \sigma^2_\alpha)$
- This formulation leads naturally into expressing the model as a regression with multiple error terms:
  $y_i = X_i\beta + \mu_\alpha + \eta_j + \sigma^2_\epsilon$
- We can also express it as a standard regression with correlated errors:
  $y_i = X_i\beta + \epsilon^{\text{all}}_i, \epsilon^{\text{all}}_i \sim N(0, \Sigma)$ where Σ is structured in a particular way.
- Generally I find it easier to think about the intercepts as latent variables, but the error formulation is more intuitive to some people.

Varying Slopes

Varying Slopes

Varying slopes are essentially the same but we now allow slope coefficients to vary, possibly via group level predictors

Varying Slopes

Varying slopes are essentially the same but we now allow slope coefficients to vary, possibly via group level predictors

$$
\begin{aligned}
y_a &= \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i \\
\alpha_j &= \gamma_0^{\alpha} + \gamma_1^{\alpha} u_j + \eta_j^{\alpha} \\
\beta_j &= \gamma_0^{\beta} + \gamma_1^{\beta} u_j + \eta_j^{\beta}
\end{aligned}
$$

Varying Slopes

Varying slopes are essentially the same but we now allow slope coefficients to vary, possibly via group level predictors

$$
\begin{aligned}
y_a &= \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i \\
\alpha_j &= \gamma_0^\alpha + \gamma_1^\alpha u_j + \eta_j^\alpha \\
\beta_j &= \gamma_0^\beta + \gamma_1^\beta u_j + \eta_j^\beta
\end{aligned}
$$

We can re-express this as a regression with interactions:

$$
y_i = \left[\gamma_0^\alpha + \gamma_1^\alpha u_{j[i]} + \eta_{j[i]}^\alpha\right] + \left[\gamma_0^\beta + \gamma_1^\beta u_{j[i]} + \eta_{j[i]}^\beta\right] x_i + \epsilon_i
$$

Varying Slopes

Varying slopes are essentially the same but we now allow slope coefficients to vary, possibly via group level predictors

$$\begin{aligned} y_a &= \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i \\ \alpha_j &= \gamma_0^\alpha + \gamma_1^\alpha u_j + \eta_j^\alpha \\ \beta_j &= \gamma_0^\beta + \gamma_1^\beta u_j + \eta_j^\beta \end{aligned}$$

We can re-express this as a regression with interactions:

$$y_i = \left[\gamma_0^\alpha + \gamma_1^\alpha u_{j[i]} + \eta_{j[i]}^\alpha\right] + \left[\gamma_0^\beta + \gamma_1^\beta u_{j[i]} + \eta_{j[i]}^\beta\right] x_i + \epsilon_i$$

Treating $u_{j[i]}$ as an individual level predictor, we can see that this is a model with interactions between x and all the group indicators, and between x and between u and x.

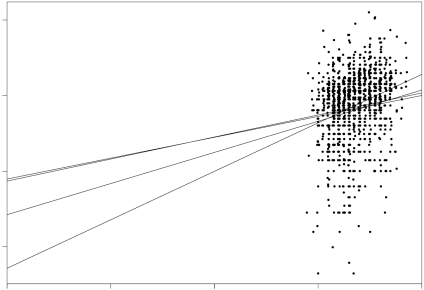
Centering

There are a lot of different perspectives on covariate centering in the literature, with more or less attention given depending on the source.

Centering

There are a lot of different perspectives on covariate centering in the literature, with more or less attention given depending on the source.

Centering can have large impacts on speed of convergence in estimation but also interpretation. The intuition for interpretation differences follows from the analog to interactions.



Distributions for Slope Models

The strong correlation between the slope and the intercept needs to be included in our model .

$$
\begin{aligned}
y_i &\sim N(X_i, \beta_{j[i]}, \sigma_y^2) \\
\beta_j &\sim N(M_B, \Sigma_B)
\end{aligned}
$$

Distributions for Slope Models

The strong correlation between the slope and the intercept needs to be included in our model .

$$y_i \sim N(X_i, \beta_{j[i]}, \sigma_y^2)$$
$$\beta_j \sim N(M_B, \Sigma_B)$$

The complexity arises in how to model Σ_B.

Distributions for Slope Models

The strong correlation between the slope and the intercept needs to be included in our model .

$$y_i \sim N(X_i, \beta_{j[i]}, \sigma_y^2)$$
$$\beta_j \sim N(M_B, \Sigma_B)$$

The complexity arises in how to model Σ_B.

Gelman recommends a scaled inverse-Wishart distribution which we won't discuss now. See Gelman and Hill (2007) Chapter 13 for more.

Additional Complexities

Additional Complexities

- A major advantage of the multilevel infrastructure is the generality

Additional Complexities

- A major advantage of the multilevel infrastructure is the generality
- Models can in theory be used for arbitrary depths and for non-nested groups

Additional Complexities

- A major advantage of the multilevel infrastructure is the generality
- Models can in theory be used for arbitrary depths and for non-nested groups
- The methods for linear models can also be extended to generalized linear models. Estimation gets harder but most other things are the same.

Additional Complexities

- A major advantage of the multilevel infrastructure is the generality
- Models can in theory be used for arbitrary depths and for non-nested groups
- The methods for linear models can also be extended to generalized linear models. Estimation gets harder but most other things are the same.
- Different types of smoothing can be imposed when groups are ordered either temporally, spatially or both

Additional Complexities

- A major advantage of the multilevel infrastructure is the generality
- Models can in theory be used for arbitrary depths and for non-nested groups
- The methods for linear models can also be extended to generalized linear models. Estimation gets harder but most other things are the same.
- Different types of smoothing can be imposed when groups are ordered either temporally, spatially or both
- Many large classes of models are simply special cases of the hierarchical models considered here

Additional Complexities

- A major advantage of the multilevel infrastructure is the generality
- Models can in theory be used for arbitrary depths and for non-nested groups
- The methods for linear models can also be extended to generalized linear models. Estimation gets harder but most other things are the same.
- Different types of smoothing can be imposed when groups are ordered either temporally, spatially or both
- Many large classes of models are simply special cases of the hierarchical models considered here
- The downside is that things get complicated quickly- which is why focused treatments of these specialized cases are important!

Estimation

- There are an enormous number of ways to fit hierarchical models and (to make matters worse!) many monikers for each strategy

Estimation

- There are an enormous number of ways to fit hierarchical models and (to make matters worse!) many monikers for each strategy
- The two most relevant strategies for our purposes are: Restricted Maximum Likelihood (REML) and Markov Chain Monte Carlo (MCMC)

Estimation

- There are an enormous number of ways to fit hierarchical models and (to make matters worse!) many monikers for each strategy
- The two most relevant strategies for our purposes are: Restricted Maximum Likelihood (REML) and Markov Chain Monte Carlo (MCMC)
- Both strategies have a number of variants and there are various ways that even using those names isn't quite right.

Estimation

- There are an enormous number of ways to fit hierarchical models and (to make matters worse!) many monikers for each strategy
- The two most relevant strategies for our purposes are: Restricted Maximum Likelihood (REML) and Markov Chain Monte Carlo (MCMC)
- Both strategies have a number of variants and there are various ways that even using those names isn't quite right.
- Let's focus on two alternatives in `R` which are both important in their own right: `lmer` in `lme4` and `rstanarm`

Estimation

- There are an enormous number of ways to fit hierarchical models and (to make matters worse!) many monikers for each strategy
- The two most relevant strategies for our purposes are: Restricted Maximum Likelihood (REML) and Markov Chain Monte Carlo (MCMC)
- Both strategies have a number of variants and there are various ways that even using those names isn't quite right.
- Let's focus on two alternatives in `R` which are both important in their own right: `lmer` in `lme4` and `rstanarm`
- Stan is a cross-platform probabilistic programming language. It can be used to expand to almost any model you can dream of.

Concluding Thoughts: What You Don't Know

- This has been a teaser for hierarchical models- there is a huge amount not covered here and blindly jumping can result in things going wrong

Concluding Thoughts: What You Don't Know

- This has been a teaser for hierarchical models- there is a huge amount not covered here and blindly jumping can result in things going wrong
- Hopefully now though you have (a) an intuition for how hierarchical models work and (b) a foundation from which to learn more.

Concluding Thoughts: What You Don't Know

- This has been a teaser for hierarchical models- there is a huge amount not covered here and blindly jumping can result in things going wrong
- Hopefully now though you have (a) an intuition for how hierarchical models work and (b) a foundation from which to learn more.
- ~~When in doubt~~ Always check your models!

Concluding Thoughts: What You Don't Know

- This has been a teaser for hierarchical models- there is a huge amount not covered here and blindly jumping can result in things going wrong
- Hopefully now though you have (a) an intuition for how hierarchical models work and (b) a foundation from which to learn more.
- ~~When in doubt~~ Always check your models!
- Read Chapter 21: "Understanding and summarizing the fitted models" in Gelman and Hill (2007)

Concluding Thoughts: What You Don't Know

- This has been a teaser for hierarchical models- there is a huge amount not covered here and blindly jumping can result in things going wrong
- Hopefully now though you have (a) an intuition for how hierarchical models work and (b) a foundation from which to learn more.
- ~~When in doubt~~ Always check your models!
- Read Chapter 21: "Understanding and summarizing the fitted models" in Gelman and Hill (2007)
- Gelman and Hill (2007) is great, but the computation has modernized a bit (due to Gelman's own work!) and you should use Stan for computation over the book recommended BUGS.