Introduction to Modern Regression and Predictive Modeling

Merlise Clyde

8/29/2018

Coordinates

- Instructor: Merlise Clyde

Coordinates

- Instructor: Merlise Clyde
- TAs:

Coordinates

- Instructor: Merlise Clyde
- TAs:
  - Jiurui Tang

Coordinates

- Instructor: Merlise Clyde
- TAs:
  - Jiurui Tang
  - Abbas Zaidi

Coordinates

- Instructor: Merlise Clyde
- TAs:
  - Jiurui Tang
  - Abbas Zaidi
- Course Websites:

Coordinates

- Instructor: Merlise Clyde
- TAs:
  - Jiurui Tang
  - Abbas Zaidi
- Course Websites:
  - Main http://stat.duke.edu/courses/Fall18/sta521

Coordinates

- Instructor: Merlise Clyde
- TAs:
  - Jiurui Tang
  - Abbas Zaidi
- Course Websites:
  - Main http://stat.duke.edu/courses/Fall18/sta521
  - Sakai https://sakai.duke.edu/portal/site/sta521-f18

Coordinates

- Instructor: Merlise Clyde
- TAs:
  - Jiurui Tang
  - Abbas Zaidi
- Course Websites:
  - Main http://stat.duke.edu/courses/Fall18/sta521
  - Sakai https://sakai.duke.edu/portal/site/sta521-f18
  - Github https://github.com/STA521-F18

Grading

Component	Percentage
Participation	5%
Homework	25%
Midterm 1	20%
Midterm 2	20%
Data Analysis Part I	15%
Data Analysis Part II	15%

Groups

- Team based data analysis assignments

Groups

- Team based data analysis assignments
  - Roughly weekly assignments

Groups

- Team based data analysis assignments
  - Roughly weekly assignments
  - 10 - 20 hours of work each

Groups

- Team based data analysis assignments
  - Roughly weekly assignments
  - 10 - 20 hours of work each
  - Peer review at the end

Groups

- Team based data analysis assignments
  - Roughly weekly assignments
  - 10 - 20 hours of work each
  - Peer review at the end
- Periodic individual assignments for concepts/theory

Groups

- Team based data analysis assignments
  - Roughly weekly assignments
  - 10 - 20 hours of work each
  - Peer review at the end
- Periodic individual assignments for concepts/theory
- Expectations and roles

Groups

- Team based data analysis assignments
  - Roughly weekly assignments
  - 10 - 20 hours of work each
  - Peer review at the end
- Periodic individual assignments for concepts/theory
- Expectations and roles
  - Everyone is expected to contribute equally

Groups

- Team based data analysis assignments
  - Roughly weekly assignments
  - 10 - 20 hours of work each
  - Peer review at the end
- Periodic individual assignments for concepts/theory
- Expectations and roles
  - Everyone is expected to contribute equally
  - Everyone is expected to understand *all* code turned in

Groups

- Team based data analysis assignments
  - Roughly weekly assignments
  - 10 - 20 hours of work each
  - Peer review at the end
- Periodic individual assignments for concepts/theory
- Expectations and roles
  - Everyone is expected to contribute equally
  - Everyone is expected to understand *all* code turned in
  - Individual contribution evaluated by peer assessment

Groups

- Team based data analysis assignments
  - Roughly weekly assignments
  - 10 - 20 hours of work each
  - Peer review at the end
- Periodic individual assignments for concepts/theory
- Expectations and roles
  - Everyone is expected to contribute equally
  - Everyone is expected to understand *all* code turned in
  - Individual contribution evaluated by peer assessment
  - You may help each other, but submitted work must be your own

Policies

- Duke Community Standard

Policies

- Duke Community Standard
  - I will not lie, cheat, or steal in my academic endeavors

Policies

- Duke Community Standard
  - I will not lie, cheat, or steal in my academic endeavors
  - I will conduct myself honorably in all of my endeavors; and

Policies

- Duke Community Standard
  - I will not lie, cheat, or steal in my academic endeavors
  - I will conduct myself honorably in all of my endeavors; and
  - I will act if the standard is compromised

Policies

- Duke Community Standard
  - I will not lie, cheat, or steal in my academic endeavors
  - I will conduct myself honorably in all of my endeavors; and
  - I will act if the standard is compromised
- Plagiarism

Policies

- Duke Community Standard
  - I will not lie, cheat, or steal in my academic endeavors
  - I will conduct myself honorably in all of my endeavors; and
  - I will act if the standard is compromised
- Plagiarism
  - Use online resources (Stack-exchange, etc) but make sure to cite them (code or theory)

Policies

- Duke Community Standard
  - I will not lie, cheat, or steal in my academic endeavors
  - I will conduct myself honorably in all of my endeavors; and
  - I will act if the standard is compromised
- Plagiarism
  - Use online resources (Stack-exchange, etc) but make sure to cite them (code or theory)
  - No direct code sharing between groups / individuals

Policies

- Duke Community Standard
  - I will not lie, cheat, or steal in my academic endeavors
  - I will conduct myself honorably in all of my endeavors; and
  - I will act if the standard is compromised
- Plagiarism
  - Use online resources (Stack-exchange, etc) but make sure to cite them (code or theory)
  - No direct code sharing between groups / individuals
- Coding Homework

Policies

- Duke Community Standard
  - I will not lie, cheat, or steal in my academic endeavors
  - I will conduct myself honorably in all of my endeavors; and
  - I will act if the standard is compromised
- Plagiarism
  - Use online resources (Stack-exchange, etc) but make sure to cite them (code or theory)
  - No direct code sharing between groups / individuals
- Coding Homework
  - Group based, everyone is equally responsible

Policies

- Duke Community Standard
  - I will not lie, cheat, or steal in my academic endeavors
  - I will conduct myself honorably in all of my endeavors; and
  - I will act if the standard is compromised
- Plagiarism
  - Use online resources (Stack-exchange, etc) but make sure to cite them (code or theory)
  - No direct code sharing between groups / individuals
- Coding Homework
  - Group based, everyone is equally responsible
- Late Homework Policy:

Policies

- Duke Community Standard
  - I will not lie, cheat, or steal in my academic endeavors
  - I will conduct myself honorably in all of my endeavors; and
  - I will act if the standard is compromised
- Plagiarism
  - Use online resources (Stack-exchange, etc) but make sure to cite them (code or theory)
  - No direct code sharing between groups / individuals
- Coding Homework
  - Group based, everyone is equally responsible
- Late Homework Policy:
  - No Late HW

Policies

- Duke Community Standard
  - I will not lie, cheat, or steal in my academic endeavors
  - I will conduct myself honorably in all of my endeavors; and
  - I will act if the standard is compromised
- Plagiarism
  - Use online resources (Stack-exchange, etc) but make sure to cite them (code or theory)
  - No direct code sharing between groups / individuals
- Coding Homework
  - Group based, everyone is equally responsible
- Late Homework Policy:
  - No Late HW
  - Drop the lowest score

Policies

- Duke Community Standard
  - I will not lie, cheat, or steal in my academic endeavors
  - I will conduct myself honorably in all of my endeavors; and
  - I will act if the standard is compromised
- Plagiarism
  - Use online resources (Stack-exchange, etc) but make sure to cite them (code or theory)
  - No direct code sharing between groups / individuals
- Coding Homework
  - Group based, everyone is equally responsible
- Late Homework Policy:
  - No Late HW
  - Drop the lowest score
- 2 In-Class Midterms

Reproducible Research / Data Analysis

- R + RStudio + JAGS

Reproducible Research / Data Analysis

- R + RStudio + JAGS
- Rmarkdown/knitr

Reproducible Research / Data Analysis

- R + RStudio + JAGS
- Rmarkdown/knitr
- Git + github

For Friday

- Install recommended software

For Friday

- Install recommended software
  - R

For Friday

- Install recommended software
  - R
  - Rstudio

For Friday

- Install recommended software
  - R
  - Rstudio
  - JAGS

For Friday

- Install recommended software
  - R
  - Rstudio
  - JAGS
- Try R Code School if you are new to R

For Friday

- Install recommended software
  - R
  - Rstudio
  - JAGS
- Try R Code School if you are new to R
- Create a github account (if you do not have one already)

For Friday

- Install recommended software
  - R
  - Rstudio
  - JAGS
- Try R Code School if you are new to R
- Create a github account (if you do not have one already)
- Complete the course survey (email link next week)

Data Science

See (Bin Yu's IMS Presidential Address 2014)[http://bulletin.imstat.org/2014/10/ims-presidential-address-let-us-own-data-science/]

Modern Regression & Predictive Modelling

- Response variable Y_i

Modern Regression & Predictive Modelling

- Response variable Y_i
- Inputs X_i (vector)

Modern Regression & Predictive Modelling

- Response variable Y_i
- Inputs X_i (vector)
- Goals:

Modern Regression & Predictive Modelling

- Response variable Y_i
- Inputs X_i (vector)
- Goals:
  - learn a model to **predict** Y_i given X_i at new inputs X_i

Modern Regression & Predictive Modelling

- Response variable Y_i
- Inputs X_i (vector)
- Goals:
  - learn a model to **predict** Y_i given X_i at new inputs X_i
  - understand relationship between X_i and Y_i (**inference**)

Modern Regression & Predictive Modelling

- Response variable Y_i
- Inputs X_i (vector)
- Goals:
  - learn a model to **predict** Y_i given X_i at new inputs X_i
  - understand relationship between X_i and Y_i (**inference**)
  - $E[Y_i] = f(X_i)$ learn regression function $f(X)$

Modern Regression & Predictive Modelling

- Response variable Y_i
- Inputs X_i (vector)
- Goals:
  - learn a model to **predict** Y_i given X_i at new inputs X_i
  - understand relationship between X_i and Y_i (**inference**)
  - $E[Y_i] = f(X_i)$ learn regression function $f(X)$
  - Model Based Statistical Learning

Course Expectations

- Expect to deal with simple to increasingly messy data (real world)

Course Expectations

- Expect to deal with simple to increasingly messy data (real world)
- Writing `R` and `JAGS` code that is reproducible

Course Expectations

- Expect to deal with simple to increasingly messy data (real world)
- Writing `R` and `JAGS` code that is reproducible
- self-documented code/solutions using `Rmarkdown`

Course Expectations

- Expect to deal with simple to increasingly messy data (real world)
- Writing `R` and `JAGS` code that is reproducible
- self-documented code/solutions using `Rmarkdown`
- use of version control (git) for team based reproducible coding

Course Expectations

- Expect to deal with simple to increasingly messy data (real world)
- Writing `R` and `JAGS` code that is reproducible
- self-documented code/solutions using `Rmarkdown`
- use of version control (git) for team based reproducible coding
- interpretation of results for non-statisticians

Course Topics

- Visualization and Exploratory Data Analysis

Course Topics

- Visualization and Exploratory Data Analysis
- Linear Regression

Course Topics

- Visualization and Exploratory Data Analysis
- Linear Regression
- Diagnostics and model checking

Course Topics

- Visualization and Exploratory Data Analysis
- Linear Regression
- Diagnostics and model checking
- Predictive Distributions

Course Topics

- Visualization and Exploratory Data Analysis
- Linear Regression
- Diagnostics and model checking
- Predictive Distributions
- Model Selection including variable selection, variable transformations, distribution choices

Course Topics

- Visualization and Exploratory Data Analysis
- Linear Regression
- Diagnostics and model checking
- Predictive Distributions
- Model Selection including variable selection, variable transformations, distribution choices
- Model Uncertainty (Bayesian Model Averaging and other Ensemble Methods)

Course Topics

- Visualization and Exploratory Data Analysis
- Linear Regression
- Diagnostics and model checking
- Predictive Distributions
- Model Selection including variable selection, variable transformations, distribution choices
- Model Uncertainty (Bayesian Model Averaging and other Ensemble Methods)
- Bayesian Shrinkage and Penalized Likelihood Estimation (Ridge Regression/ LASSO/ Horseshoe)

Course Topics

- Visualization and Exploratory Data Analysis
- Linear Regression
- Diagnostics and model checking
- Predictive Distributions
- Model Selection including variable selection, variable transformations, distribution choices
- Model Uncertainty (Bayesian Model Averaging and other Ensemble Methods)
- Bayesian Shrinkage and Penalized Likelihood Estimation (Ridge Regression/ LASSO/ Horseshoe)
- Robust Estimation

Course Topics

- Visualization and Exploratory Data Analysis
- Linear Regression
- Diagnostics and model checking
- Predictive Distributions
- Model Selection including variable selection, variable transformations, distribution choices
- Model Uncertainty (Bayesian Model Averaging and other Ensemble Methods)
- Bayesian Shrinkage and Penalized Likelihood Estimation (Ridge Regression/ LASSO/ Horseshoe)
- Robust Estimation
- Classification and Regression Trees, Random Forests, Boosting, Bayesian Additive Regression Trees

Course Topics

- Visualization and Exploratory Data Analysis
- Linear Regression
- Diagnostics and model checking
- Predictive Distributions
- Model Selection including variable selection, variable transformations, distribution choices
- Model Uncertainty (Bayesian Model Averaging and other Ensemble Methods)
- Bayesian Shrinkage and Penalized Likelihood Estimation (Ridge Regression/ LASSO/ Horseshoe)
- Robust Estimation
- Classification and Regression Trees, Random Forests, Boosting, Bayesian Additive Regression Trees
- Other Topics: Nonparametric Regression, Time Series, Neural Networks

Themes

- Interpretability versus predictive performance

Themes

- Interpretability versus predictive performance
- Bias-Variance Trade-off

Themes

- Interpretability versus predictive performance
- Bias-Variance Trade-off
- In sample versus out-of-sample

Themes

- Interpretability versus predictive performance
- Bias-Variance Trade-off
- In sample versus out-of-sample
- point estimates versus uncertainty quantification

Themes

- Interpretability versus predictive performance
- Bias-Variance Trade-off
- In sample versus out-of-sample
- point estimates versus uncertainty quantification
- exact analysis versus approximation (computational scaling)

Themes

- Interpretability versus predictive performance
- Bias-Variance Trade-off
- In sample versus out-of-sample
- point estimates versus uncertainty quantification
- exact analysis versus approximation (computational scaling)
- understanding structure of data (relationships)

Themes

- Interpretability versus predictive performance
- Bias-Variance Trade-off
- In sample versus out-of-sample
- point estimates versus uncertainty quantification
- exact analysis versus approximation (computational scaling)
- understanding structure of data (relationships)
- Bayesian versus Frequentist ?

Tradeoffs...

Frequenstist & Bayes

- Likelihood Based inference

Frequenstist & Bayes

- Likelihood Based inference
  - Sampling model

$$Y_i \sim f(y_i \mid \theta)$$

Frequenstist & Bayes

- Likelihood Based inference
  - Sampling model
    $$Y_i \sim f(y_i \mid \theta)$$
  - Likelihood
    $$L(\theta) \propto \prod_i f(Y_i \mid \theta)$$

Frequenstist & Bayes

- Likelihood Based inference
  - Sampling model
$$Y_i \sim f(y_i \mid \theta)$$
  - Likelihood
$$L(\theta) \propto \prod_i f(Y_i \mid \theta)$$
  - MLE of θ

Frequenstist & Bayes

- Likelihood Based inference
  - Sampling model
$$Y_i \sim f(y_i \mid \theta)$$
  - Likelihood
$$L(\theta) \propto \prod_i f(Y_i \mid \theta)$$
  - MLE of θ
- Bayes

Frequenstist & Bayes

- Likelihood Based inference
  - Sampling model
    $$Y_i \sim f(y_i \mid \theta)$$
  - Likelihood
    $$L(\theta) \propto \prod_i f(Y_i \mid \theta)$$
  - MLE of θ
- Bayes
  - prior distribution $p(\theta)$ describes prior uncertainty about θ

Frequenstist & Bayes

- Likelihood Based inference
  - Sampling model
    $$Y_i \sim f(y_i \mid \theta)$$
  - Likelihood
    $$L(\theta) \propto \prod_i f(Y_i \mid \theta)$$
  - MLE of θ
- Bayes
  - prior distribution $p(\theta)$ describes prior uncertainty about θ
  - posterior distribution

$$p(\theta \mid data) \propto L(\theta)p(\theta)$$

Frequenstist & Bayes

- Likelihood Based inference
  - Sampling model
$$Y_i \sim f(y_i \mid \theta)$$
  - Likelihood
$$L(\theta) \propto \prod_i f(Y_i \mid \theta)$$
  - MLE of θ
- Bayes
  - prior distribution $p(\theta)$ describes prior uncertainty about θ
  - posterior distribution
$$p(\theta \mid data) \propto L(\theta)p(\theta)$$
  - uncertainty after seeing data

Got Data?

&h=800&hash=sfmK8PW%2BTbPHupf8ExHOszMCVRg%3D&ora=1%

800&hash =

sfmK8PW%2BTbPHupf8ExHOszMCVRg%3D&ora =

1%2CaFBCTXdkRmpGL2lvQUFBPQ%2CxAVta9Er0Vinkhwfjw

800&hash =

sfmK8PW%2BTbPHupf8ExHOszMCVRg%3D&ora =

1%2CaFBCTXdkRmpGL2lvQUFBPQ%2CxAVta9Er0Vinkhwfjw

800&hash =

sfmK8PW%2BTbPHupf8ExHOszMCVRg%3D&ora =

1%2CaFBCTXdkRmpGL2lvQUFBPQ%2CxAVta9Er0Vinkhwfjw

800&hash =

sfmK8PW%2BTbPHupf8ExHOszMCVRg%3D&ora =

1%2CaFBCTXdkRmpGL2lvQUFBPQ%2CxAVta9Er0Vinkhwfjw

800&hash =

sfmK8PW%2BTbPHupf8ExHOszMCVRg%3D&ora =

1%2CaFBCTXdkRmpGL2lvQUFBPQ%2CxAVta9Er0Vinkhwfjw

800&hash =

sfmK8PW%2BTbPHupf8ExHOszMCVRg%3D&ora =

Philosophy

- for many problems Frequentist and Bayesian methods will give similar answers (more a matter of taste in interpretation)

Philosophy

- for many problems Frequentist and Bayesian methods will give similar answers (more a matter of taste in interpretation)
  - For small problems, Bayesian methods allow us to incorporate prior information which provides better calibrated answers and better measure of uncertainty

Philosophy

- for many problems Frequentist and Bayesian methods will give similar answers (more a matter of taste in interpretation)
  - For small problems, Bayesian methods allow us to incorporate prior information which provides better calibrated answers and better measure of uncertainty
  - for problems with complex designs and/or missing data Bayesian methods are often better easier to implement (do not need to rely on asymptotics)

Philosophy

- for many problems Frequentist and Bayesian methods will give similar answers (more a matter of taste in interpretation)
  - For small problems, Bayesian methods allow us to incorporate prior information which provides better calibrated answers and better measure of uncertainty
  - for problems with complex designs and/or missing data Bayesian methods are often better easier to implement (do not need to rely on asymptotics)
- For problems involving hypothesis testing or model selection Frequentist and Bayesian methods can be strikingly different.

Philosophy

- for many problems Frequentist and Bayesian methods will give similar answers (more a matter of taste in interpretation)
  - For small problems, Bayesian methods allow us to incorporate prior information which provides better calibrated answers and better measure of uncertainty
  - for problems with complex designs and/or missing data Bayesian methods are often better easier to implement (do not need to rely on asymptotics)
- For problems involving hypothesis testing or model selection Frequentist and Bayesian methods can be strikingly different.
- Frequentist methods often faster (particularly with "big data") so great for exploratory analysis and for building a *data-sense*

Philosophy

- for many problems Frequentist and Bayesian methods will give similar answers (more a matter of taste in interpretation)
  - For small problems, Bayesian methods allow us to incorporate prior information which provides better calibrated answers and better measure of uncertainty
  - for problems with complex designs and/or missing data Bayesian methods are often better easier to implement (do not need to rely on asymptotics)
- For problems involving hypothesis testing or model selection Frequentist and Bayesian methods can be strikingly different.
- Frequentist methods often faster (particularly with "big data") so great for exploratory analysis and for building a *data-sense*
- Bayesian methods sit on top of Frequentist Likelihood

Philosophy

- for many problems Frequentist and Bayesian methods will give similar answers (more a matter of taste in interpretation)
  - For small problems, Bayesian methods allow us to incorporate prior information which provides better calibrated answers and better measure of uncertainty
  - for problems with complex designs and/or missing data Bayesian methods are often better easier to implement (do not need to rely on asymptotics)
- For problems involving hypothesis testing or model selection Frequentist and Bayesian methods can be strikingly different.
- Frequentist methods often faster (particularly with "big data") so great for exploratory analysis and for building a *data-sense*
- Bayesian methods sit on top of Frequentist Likelihood
- Important to understand advantages and problems of each perspective!

Statistical and Machine Learning ?

Ovarian Cancer Risk Prediction

- Binary Outcome (Cancer/Control)

Ovarian Cancer Risk Prediction

- Binary Outcome (Cancer/Control)
  - 17 established SNPS (genetic markers)

Ovarian Cancer Risk Prediction

- Binary Outcome (Cancer/Control)
  - 17 established SNPS (genetic markers)
  - other risk factors (age, family history, oral contraceptive use, number of pregnancies, aspirin, ...)

Ovarian Cancer Risk Prediction

- Binary Outcome (Cancer/Control)
  - 17 established SNPS (genetic markers)
  - other risk factors (age, family history, oral contraceptive use, number of pregnancies, aspirin, ...)
- Case - Control design

Ovarian Cancer Risk Prediction

- Binary Outcome (Cancer/Control)
  - 17 established SNPS (genetic markers)
  - other risk factors (age, family history, oral contraceptive use, number of pregnancies, aspirin, ...)
- Case - Control design
- variability across study sites (random effects)

Ovarian Cancer Risk Prediction

- Binary Outcome (Cancer/Control)
  - 17 established SNPS (genetic markers)
  - other risk factors (age, family history, oral contraceptive use, number of pregnancies, aspirin, ...)
- Case - Control design
- variability across study sites (random effects)
- 80% subjects had at least one variable with missing data

Ovarian Cancer Risk Prediction

- Binary Outcome (Cancer/Control)
  - 17 established SNPS (genetic markers)
  - other risk factors (age, family history, oral contraceptive use, number of pregnancies, aspirin, ...)
- Case - Control design
- variability across study sites (random effects)
- 80% subjects had at least one variable with missing data
- Missing at random versus missing not-at-random

Ovarian Cancer Risk Prediction

- Binary Outcome (Cancer/Control)
  - 17 established SNPS (genetic markers)
  - other risk factors (age, family history, oral contraceptive use, number of pregnancies, aspirin, ...)
- Case - Control design
- variability across study sites (random effects)
- 80% subjects had at least one variable with missing data
- Missing at random versus missing not-at-random
- Focus is on prediction, but still need an interpretable model

Ovarian Cancer Risk Prediction

- Binary Outcome (Cancer/Control)
  - 17 established SNPS (genetic markers)
  - other risk factors (age, family history, oral contraceptive use, number of pregnancies, aspirin, ...)
- Case - Control design
- variability across study sites (random effects)
- 80% subjects had at least one variable with missing data
- Missing at random versus missing not-at-random
- Focus is on prediction, but still need an interpretable model
- EDA, Model Building, and Predictive Checking crucial!