Arsenic and old models

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

6 Sept 2007

Contents

Contents

- Aresenic in Bangladesh
- Decision analysis
- Regression models

Contents

Contents

- Aresenic in Bangladesh
- Decision analysis
- Regression models

Natural arsenic in well water in Bangladesh

- Where is the arsenic?
- What can people do?
- Digging low-arsenic wells
- Will people switch?

Natural arsenic in well water in Bangladesh

- Where is the arsenic?
- What can people do?
- Digging low-arsenic wells
- Will people switch?

Natural arsenic in well water in Bangladesh

- Where is the arsenic?
- What can people do?
- Digging low-arsenic wells
- Will people switch?

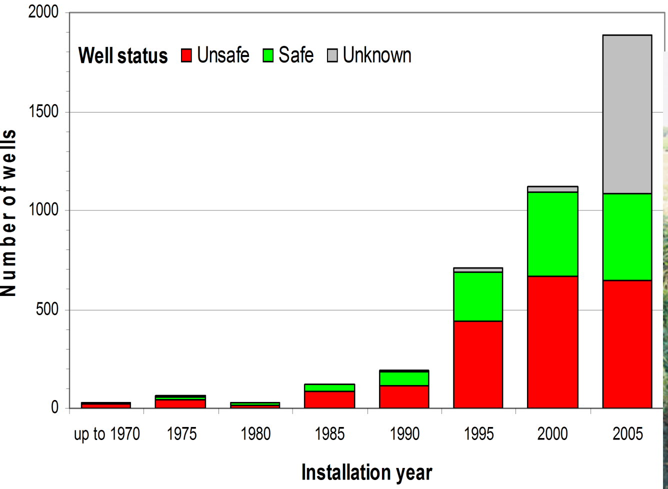
Natural arsenic in well water in Bangladesh

- Where is the arsenic?
- What can people do?
- Digging low-arsenic wells
- Will people switch?

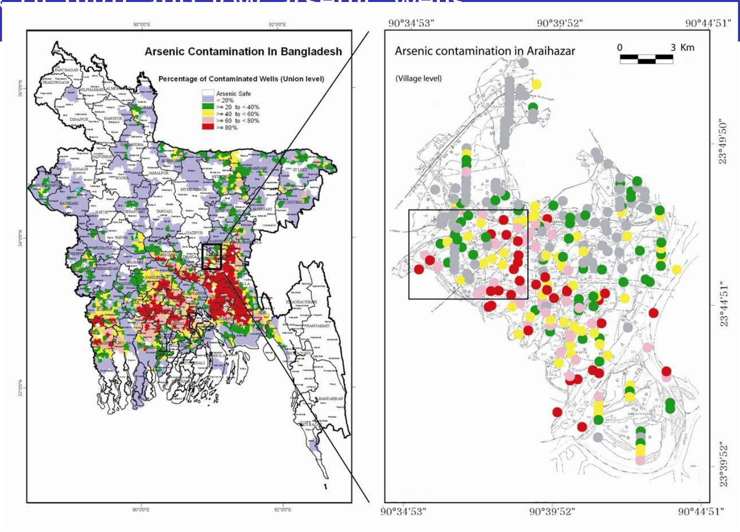
Natural arsenic in well water in Bangladesh

- Where is the arsenic?
- What can people do?
- Digging low-arsenic wells
- Will people switch?

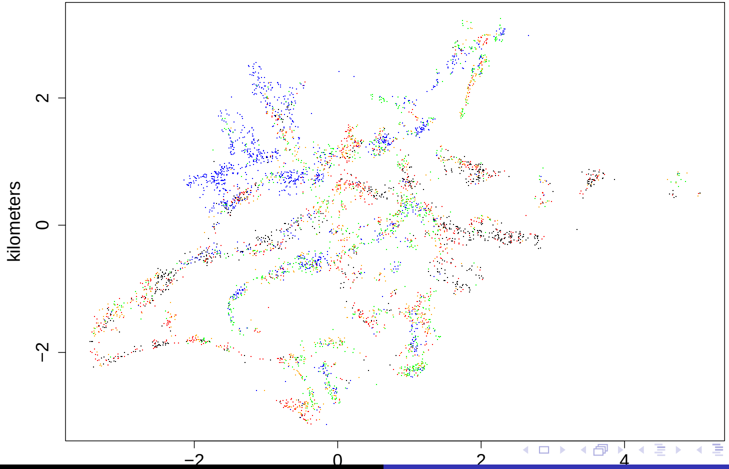
Natural arsenic in well water



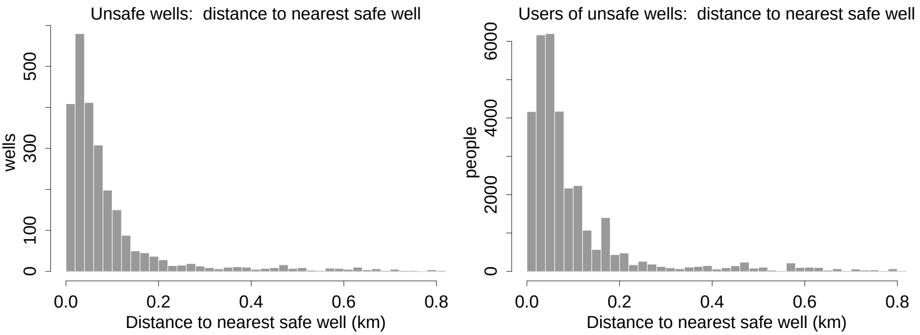
Mix of high and low arsenic wells



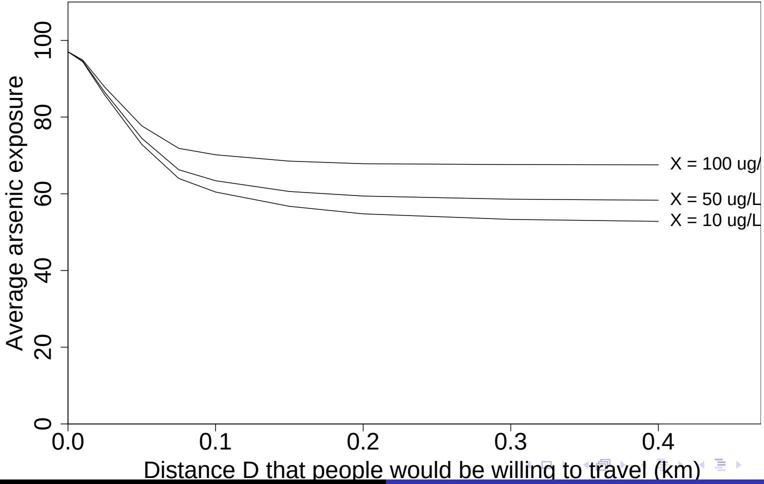
Mix of high and low arsenic wells



Distance to nearest safe well

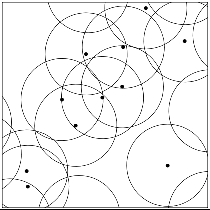


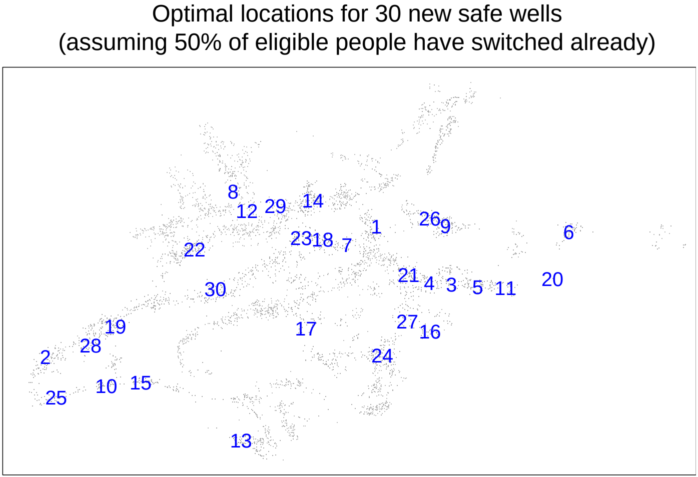
What if people switch wells?



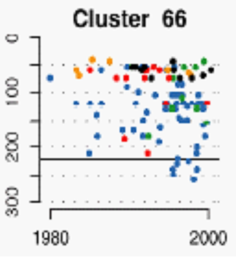
Digging new wells

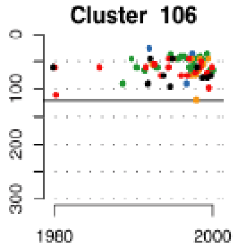
Where to dig new wells

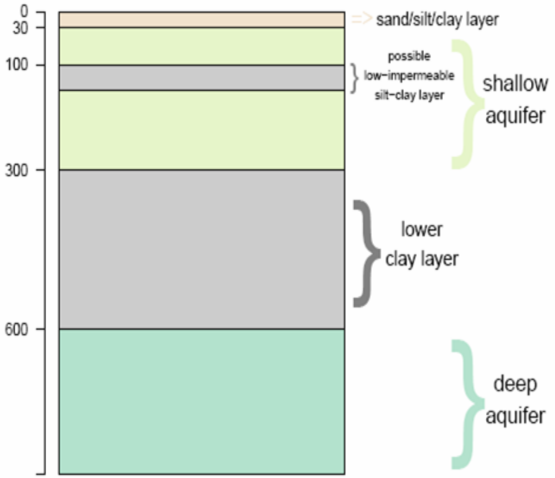




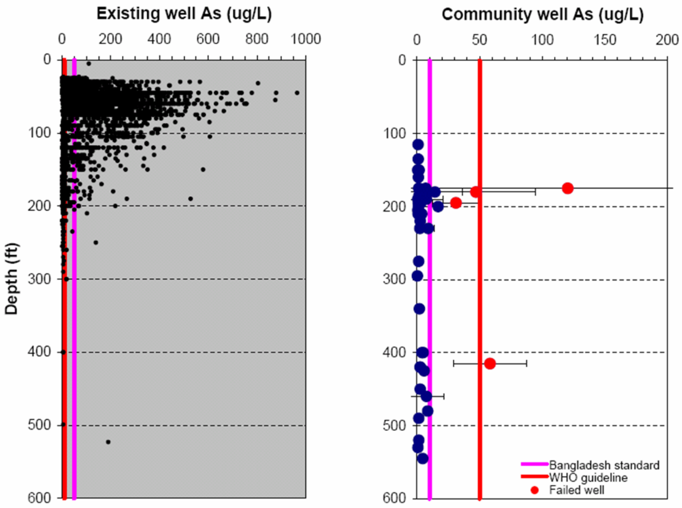
How deep to dig?



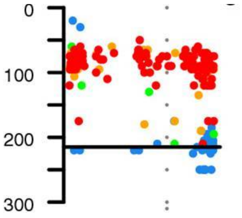




New community wells



Cellphone-based information system



Instructions:	SMS "?" to +880 1713 045 512 or http://www.ldeo.columbia.edu/welltracker/
Find village?	"F*U*Araihazar*V*Bara Barai Para"
Response:	U:Araihazar M: Bara Barai Para Bara Barai Para, 167029410201" and others
Safe depth?	"SD*167029410201"
Response:	U=Araihazar M=Bara Barai Para V=Bara Barai Para Start>=215' Fail=5/100 Average arsenic 168 ppb 39 safe of 183 20-135' 7 of 142 175-250' 32 of 41
Money-back?	"SD*167029410201*15000"
Response:	TK750 insures TK15000 (US$250) Add TK1000-3000 fixed cost (well design, test)

Survey data: would you switch wells?

- Logistic regression
- Predictor variables:
  - Distance to nearest safe well
  - Arsenic level of your current well
  - Education
  - Membership in community organizations (not predictive)

Survey data: would you switch wells?

- Logistic regression
- Predictor variables:
  - Distance to nearest safe well
  - Arsenic level of your current well
  - Education
  - Membership in community organizations (not predictive)

Survey data: would you switch wells?

- Logistic regression
- Predictor variables:
  - Distance to nearest safe well
  - Arsenic level of your current well
  - Education
  - Membership in community organizations (not predictive)

Survey data: would you switch wells?

- Logistic regression
- Predictor variables:
  - Distance to nearest safe well
  - Arsenic level of your current well
  - Education
  - Membership in community organizations (not predictive)

Survey data: would you switch wells?

- Logistic regression
- Predictor variables:
  - Distance to nearest safe well
  - Arsenic level of your current well
  - Education
  - Membership in community organizations (not predictive)

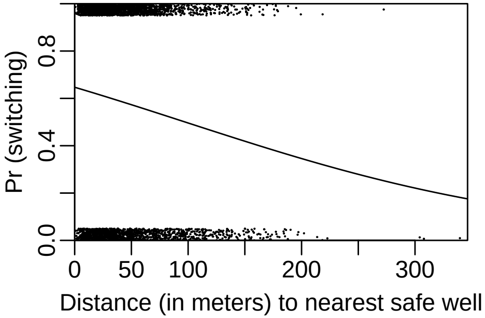
Survey data: would you switch wells?

- Logistic regression
- Predictor variables:
  - Distance to nearest safe well
  - Arsenic level of your current well
  - Education
  - Membership in community organizations (not predictive)

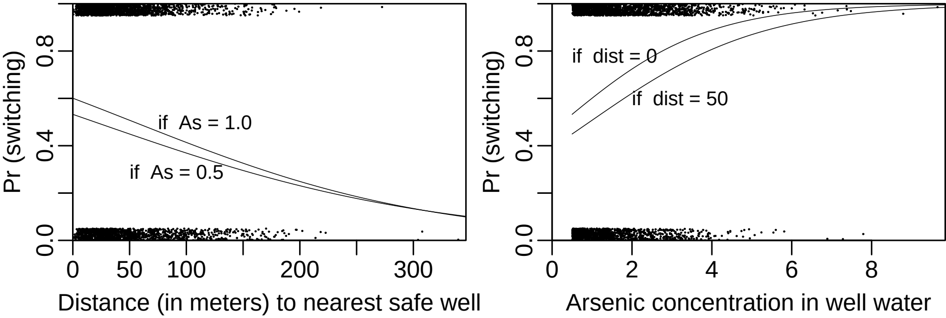
Survey data: would you switch wells?

- Logistic regression
- Predictor variables:
  - Distance to nearest safe well
  - Arsenic level of your current well
  - Education
  - Membership in community organizations (not predictive)

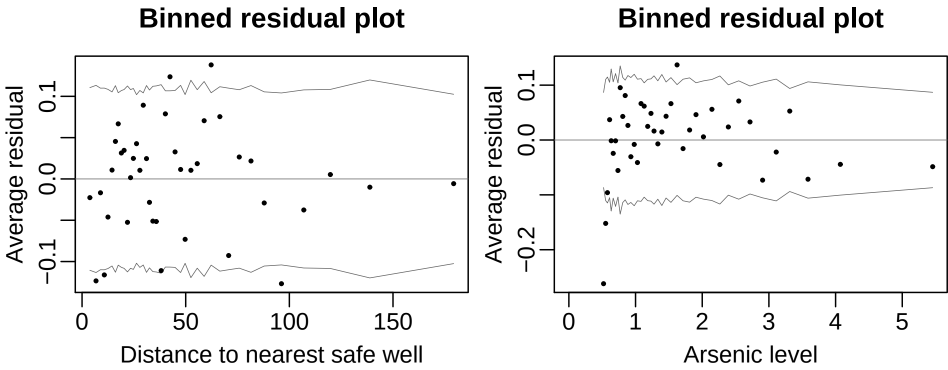
Probability of switching wells, given distance to nearest safe well



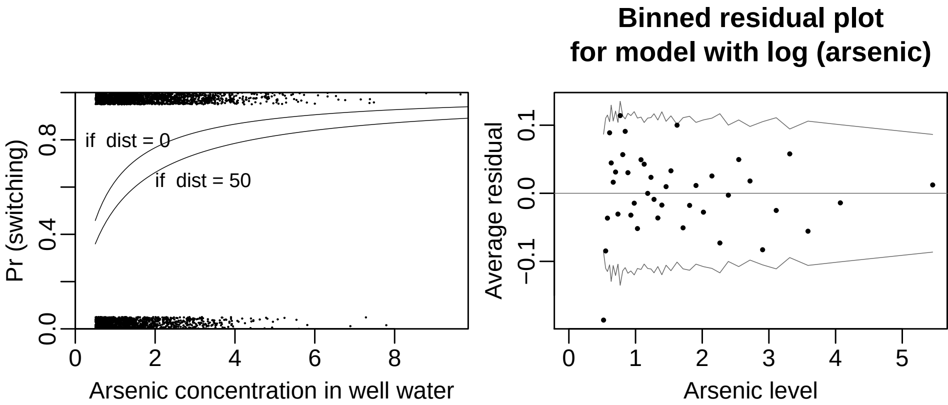
Probability of switching wells, given distance and existing arsenic level



Binned residuals: are people switching more or less than predicted by the model?



Model on log (arsenic level) and binned residuals



Model for switching

- Distance to walk comes in linearly
  - Does this make sense?
  - Yes
- Current arsenic level comes in on the log scale

Model for switching

- Distance to walk comes in linearly
  - Does this make sense?
  - Yes
- Current arsenic level comes in on the log scale

Model for switching

- Distance to walk comes in linearly
  - Does this make sense?
  - Yes
- Current arsenic level comes in on the log scale
  - Does this make sense?
  - Yes and no

Model for switching

- Distance to walk comes in linearly
  - Does this make sense?
  - Yes
- Current arsenic level comes in on the log scale
  - Does this make sense?
  - Yes and no

Model for switching

- Distance to walk comes in linearly
  - Does this make sense?
  - Yes
- Current arsenic level comes in on the log scale
  - Does this make sense?
  - Yes and no

Model for switching

- Distance to walk comes in linearly
  - Does this make sense?
  - Yes
- Current arsenic level comes in on the log scale
  - Does this make sense?
  - Yes and no

Model for switching

- Distance to walk comes in linearly
  - Does this make sense?
  - Yes
- Current arsenic level comes in on the log scale
  - Does this make sense?
  - Yes and no

Insurance program

- Goal: dig more safe wells
- Outcomes to avoid:
  - Digging an unsafe well and not testing it
  - Not digging a new well because afraid of wasting money on an unsafe well
  - Digging too shallow (risk of unsafe)
  - Digging too deep (waste of money)
- A (possible) solution: insurance or money-back guarantee

Insurance program

- Goal: dig more safe wells
- Outcomes to avoid:
  - Digging an unsafe well and not testing it
  - Not digging a new well because afraid of wasting money on an unsafe well
  - Digging too shallow (risk of unsafe)
  - Digging too deep (waste of money)
- A (possible) solution: insurance or money-back guarantee

Insurance program

- Goal: dig more safe wells
- Outcomes to avoid:
  - Digging an unsafe well and not testing it
  - Not digging a new well because afraid of wasting money on an unsafe well
  - Digging too shallow (risk of unsafe)
  - Digging too deep (waste of money)
- A (possible) solution: insurance or money-back guarantee

Insurance program

- Goal: dig more safe wells
- Outcomes to avoid:
  - Digging an unsafe well and not testing it
  - Not digging a new well because afraid of wasting money on an unsafe well
  - Digging too shallow (risk of unsafe)
  - Digging too deep (waste of money)
- A (possible) solution: insurance or money-back guarantee

Insurance program

- Goal: dig more safe wells
- Outcomes to avoid:
  - Digging an unsafe well and not testing it
  - Not digging a new well because afraid of wasting money on an unsafe well
  - Digging too shallow (risk of unsafe)
  - Digging too deep (waste of money)
- A (possible) solution: insurance or money-back guarantee

Insurance program

- Goal: dig more safe wells
- Outcomes to avoid:
  - Digging an unsafe well and not testing it
  - Not digging a new well because afraid of wasting money on an unsafe well
  - Digging too shallow (risk of unsafe)
  - Digging too deep (waste of money)
- A (possible) solution: insurance or money-back guarantee

Insurance program

- Goal: dig more safe wells
- Outcomes to avoid:
  - Digging an unsafe well and not testing it
  - Not digging a new well because afraid of wasting money on an unsafe well
  - Digging too shallow (risk of unsafe)
  - Digging too deep (waste of money)
- A (possible) solution: insurance or money-back guarantee

Insurance program

- Goal: dig more safe wells
- Outcomes to avoid:
  - Digging an unsafe well and not testing it
  - Not digging a new well because afraid of wasting money on an unsafe well
  - Digging too shallow (risk of unsafe)
  - Digging too deep (waste of money)
- A (possible) solution: insurance or money-back guarantee

Decision analysis and the garbage-in, garbage-out problem

- Radon example
- Arsenic example
- Institutional decision analysis and the role of centralized information collection and analysis

Decision analysis and the garbage-in, garbage-out problem

- Radon example
- Arsenic example
- Institutional decision analysis and the role of centralized information collection and analysis

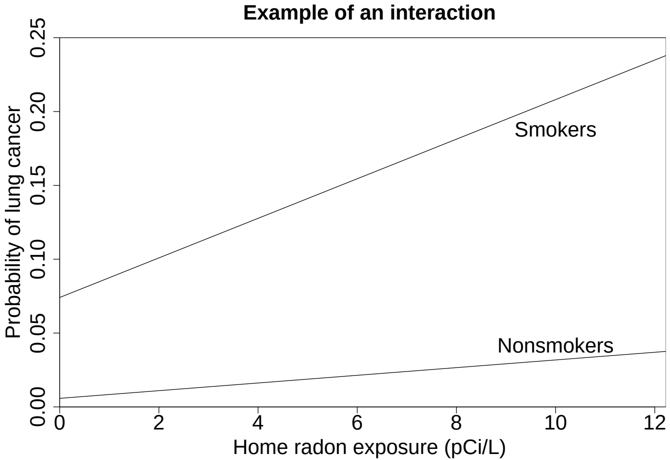
Decision analysis and the garbage-in, garbage-out problem

- Radon example
- Arsenic example
- Institutional decision analysis and the role of centralized information collection and analysis

Decision analysis and the garbage-in, garbage-out problem

- Radon example
- Arsenic example
- Institutional decision analysis and the role of centralized information collection and analysis

Radon and lung cancer: estimated risks



Home radon exposure as a decision problem

- For your house, decision options:
  - Remediate (seal the basement, etc.), costs \$2000
  - Take a good measurement, costs \$50 + wait 1 year
  - Take a noisy measurement, costs \$15 + wait 2 days
  - Do nothing
- It's a classical "value of information" decision problem!

Home radon exposure as a decision problem

- For your house, decision options:
  - Remediate (seal the basement, etc.), costs \$2000
  - Take a good measurement, costs \$50 + wait 1 year
  - Take a noisy measurement, costs \$25 + wait 1 week
  - Do nothing
- It's a classical "value of information" decision problem!

Home radon exposure as a decision problem

- For your house, decision options:
  - Remediate (seal the basement, etc.), costs \$2000
  - Take a good measurement, costs \$50 + wait 1 year
  - Take a noisy measurement, costs \$25 + wait 1 week
  - Do nothing
- It's a classical "value of information" decision problem!

Home radon exposure as a decision problem

- For your house, decision options:
  - Remediate (seal the basement, etc.), costs \$2000
  - Take a good measurement, costs \$50 + wait 1 year
  - Take a noisy measurement, costs \$25 + wait 1 week
  - Do nothing
- It's a classical "value of information" decision problem!

Home radon exposure as a decision problem

- For your house, decision options:
  - Remediate (seal the basement, etc.), costs \$2000
  - Take a good measurement, costs \$50 + wait 1 year
  - Take a noisy measurement, costs \$25 + wait 1 week
  - Do nothing
- It's a classical "value of information" decision problem!

Home radon exposure as a decision problem

- For your house, decision options:
  - Remediate (seal the basement, etc.), costs \$2000
  - Take a good measurement, costs \$50 + wait 1 year
  - Take a noisy measurement, costs \$25 + wait 1 week
  - Do nothing
- It's a classical "value of information" decision problem!

Home radon exposure as a decision problem

- For your house, decision options:
  - Remediate (seal the basement, etc.), costs \$2000
  - Take a good measurement, costs \$50 + wait 1 year
  - Take a noisy measurement, costs \$25 + wait 1 week
  - Do nothing
- It's a classical "value of information" decision problem!

Home radon analysis

- 50,000 homes with very high radon, millions with high radon
- Goal: to identify the dangerous homes
- 3 sources of information:

Home radon analysis

- 50,000 homes with very high radon, millions with high radon
- Goal: to identify the dangerous homes
- 3 sources of information:
  - National survey: accurate measurements in 5000 homes in 125 U.S. counties
  - State surveys: noisy, biased measurements in 80,000 homes in all the counties
  - County-level soil uranium measurements (from 1950s)
  - County-level geological info

Home radon analysis

- 50,000 homes with very high radon, millions with high radon
- Goal: to identify the dangerous homes
- 3 sources of information:
  - National survey: accurate measurements in 5000 homes in 125 U.S. counties
  - State surveys: noisy, biased measurements in 80,000 homes in all the counties
  - County-level soil uranium measurements (from 1950s)
  - County-level geological info

Home radon analysis

- 50,000 homes with very high radon, millions with high radon
- Goal: to identify the dangerous homes
- 3 sources of information:
  - National survey: accurate measurements in 5000 homes in 125 U.S. counties
  - State surveys: noisy, biased measurements in 80,000 homes in all the counties
  - County-level soil uranium measurements (from 1950s)
  - County-level geological info

Home radon analysis

- 50,000 homes with very high radon, millions with high radon
- Goal: to identify the dangerous homes
- 3 sources of information:
  - National survey: accurate measurements in 5000 homes in 125 U.S. counties
  - State surveys: noisy, biased measurements in 80,000 homes in all the counties
  - County-level soil uranium measurements (from 1950s)
  - County-level geological info

Home radon analysis

- 50,000 homes with very high radon, millions with high radon
- Goal: to identify the dangerous homes
- 3 sources of information:
  - National survey: accurate measurements in 5000 homes in 125 U.S. counties
  - State surveys: noisy, biased measurements in 80,000 homes in all the counties
  - County-level soil uranium measurements (from 1950s)
  - County-level geological info

Home radon analysis

- 50,000 homes with very high radon, millions with high radon
- Goal: to identify the dangerous homes
- 3 sources of information:
  - National survey: accurate measurements in 5000 homes in 125 U.S. counties
  - State surveys: noisy, biased measurements in 80,000 homes in all the counties
  - County-level soil uranium measurements (from 1950s)
  - County-level geological info

Home radon analysis

- 50,000 homes with very high radon, millions with high radon
- Goal: to identify the dangerous homes
- 3 sources of information:
  - National survey: accurate measurements in 5000 homes in 125 U.S. counties
  - State surveys: noisy, biased measurements in 80,000 homes in all the counties
  - County-level soil uranium measurements (from 1950s)
  - County-level geological info

Home radon analysis: statistical methods

- Classical method 1: use national survey to predict radon from house-level predictors (basement, ventilation, construction, county uranium, soil type, ...)
- Classical method 2: use state surveys to identify high-radon areas, them link these to geological maps
- Bayesian method: combine all the info to get inference for houses with and without basements in all counties
- Measurement-error model adjusts for low-quality data
- Cross-validation demonstrates that it works

Home radon analysis: statistical methods

- Classical method 1: use national survey to predict radon from house-level predictors (basement, ventilation, construction, county uranium, soil type, . . .)
- Classical method 2: use state surveys to identify high-radon areas, them link these to geological maps
- Bayesian method: combine all the info to get inference for houses with and without basements in all counties
- Measurement-error model adjusts for low-quality data
- Cross-validation demonstrates that it works

Home radon analysis: statistical methods

- Classical method 1: use national survey to predict radon from house-level predictors (basement, ventilation, construction, county uranium, soil type, . . .)
- Classical method 2: use state surveys to identify high-radon areas, them link these to geological maps
- Bayesian method: combine all the info to get inference for houses with and without basements in all counties
- Measurement-error model adjusts for low-quality data
- Cross-validation demonstrates that it works

Home radon analysis: statistical methods

- Classical method 1: use national survey to predict radon from house-level predictors (basement, ventilation, construction, county uranium, soil type, . . .)
- Classical method 2: use state surveys to identify high-radon areas, them link these to geological maps
- Bayesian method: combine all the info to get inference for houses with and without basements in all counties
- Measurement-error model adjusts for low-quality data
- Cross-validation demonstrates that it works

Home radon analysis: statistical methods

- Classical method 1: use national survey to predict radon from house-level predictors (basement, ventilation, construction, county uranium, soil type, . . .)
- Classical method 2: use state surveys to identify high-radon areas, them link these to geological maps
- Bayesian method: combine all the info to get inference for houses with and without basements in all counties
- Measurement-error model adjusts for low-quality data
- Cross-validation demonstrates that it works

Home radon analysis: statistical methods

- Classical method 1: use national survey to predict radon from house-level predictors (basement, ventilation, construction, county uranium, soil type, ...)
- Classical method 2: use state surveys to identify high-radon areas, them link these to geological maps
- Bayesian method: combine all the info to get inference for houses with and without basements in all counties
- Measurement-error model adjusts for low-quality data
- Cross-validation demonstrates that it works

Home radon analysis: garbage in, garbage out

- Use Bayes posterior distribution to figure out optimal decision for houses in every county
- Average over the 3000 counties to estimate the total dollar cost and lives saved under various strategies
- Compare to costs of other safety measures
- Garbage-in, garbage-out issue:
- What should the EPA say?

Home radon analysis: garbage in, garbage out

- Use Bayes posterior distribution to figure out optimal decision for houses in every county
- Average over the 3000 counties to estimate the total dollar cost and lives saved under various strategies
- Compare to costs of other safety measures
- Garbage-in, garbage-out issue:
- What should the EPA say?

Home radon analysis: garbage in, garbage out

- Use Bayes posterior distribution to figure out optimal decision for houses in every county
- Average over the 3000 counties to estimate the total dollar cost and lives saved under various strategies
- Compare to costs of other safety measures
- Garbage-in, garbage-out issue:
  - Specify your "value of a microlife" (how much you would spend to reduce risk by 1/million), or
    Specify your "action level" (the radon level at which you would do something)
- What should the EPA say?

Home radon analysis: garbage in, garbage out

- Use Bayes posterior distribution to figure out optimal decision for houses in every county
- Average over the 3000 counties to estimate the total dollar cost and lives saved under various strategies
- Compare to costs of other safety measures
- Garbage-in, garbage-out issue:
  - Specify your "value of a microlife" (how much you would spend to reduce risk by 1/million), or
  - Specify your "action level" (the radon level at which you would do something)
- What should the EPA say?

Home radon analysis: garbage in, garbage out

- Use Bayes posterior distribution to figure out optimal decision for houses in every county
- Average over the 3000 counties to estimate the total dollar cost and lives saved under various strategies
- Compare to costs of other safety measures
- Garbage-in, garbage-out issue:
  - Specify your "value of a microlife" (how much you would spend to reduce risk by 1/million), or
  - Specify your "action level" (the radon level at which you would do something)
- What should the EPA say?

Home radon analysis: garbage in, garbage out

- Use Bayes posterior distribution to figure out optimal decision for houses in every county
- Average over the 3000 counties to estimate the total dollar cost and lives saved under various strategies
- Compare to costs of other safety measures
- Garbage-in, garbage-out issue:
  - Specify your "value of a microlife" (how much you would spend to reduce risk by 1/million), or
  - Specify your "action level" (the radon level at which you would do something)
- What should the EPA say?

Home radon analysis: garbage in, garbage out

- Use Bayes posterior distribution to figure out optimal decision for houses in every county
- Average over the 3000 counties to estimate the total dollar cost and lives saved under various strategies
- Compare to costs of other safety measures
- Garbage-in, garbage-out issue:
  - Specify your "value of a microlife" (how much you would spend to reduce risk by 1/million), or
  - Specify your "action level" (the radon level at which you would do something)
- What should the EPA say?

Home radon analysis: garbage in, garbage out

- Use Bayes posterior distribution to figure out optimal decision for houses in every county
- Average over the 3000 counties to estimate the total dollar cost and lives saved under various strategies
- Compare to costs of other safety measures
- Garbage-in, garbage-out issue:
  - Specify your "value of a microlife" (how much you would spend to reduce risk by 1/million), or
  - Specify your "action level" (the radon level at which you would do something)
- What should the EPA say?

Why are textbook examples of decision analysis so lame?

- Your nephew is renting an apartment, balancing issues of cost, size, convenience, . . .
- Widgets cost $2 to make and sell for $3. Here's the distribution of the market for widgets, . . . , how many should you make?
- Vague business example
- Specific business example—what kind of power plant to build—pure GIGO
- Vague military example

Why are textbook examples of decision analysis so lame?

- Your nephew is renting an apartment, balancing issues of cost, size, convenience, . . .
- Widgets cost $2 to make and sell for $3. Here's the distribution of the market for widgets, . . . , how many should you make?
- Vague business example
- Specific business example—what kind of power plant to build—pure GIGO
- Vague military example

Why are textbook examples of decision analysis so lame?

- Your nephew is renting an apartment, balancing issues of cost, size, convenience, . . .
- Widgets cost \$2 to make and sell for \$3. Here's the distribution of the market for widgets, . . . , how many should you make?
- Vague business example
- Specific business example—what kind of power plant to build—pure GIGO
- Vague military example

Why are textbook examples of decision analysis so lame?

- Your nephew is renting an apartment, balancing issues of cost, size, convenience, . . .
- Widgets cost \$2 to make and sell for \$3. Here's the distribution of the market for widgets, . . . , how many should you make?
- Vague business example
- Specific business example—what kind of power plant to build—pure GIGO
- Vague military example

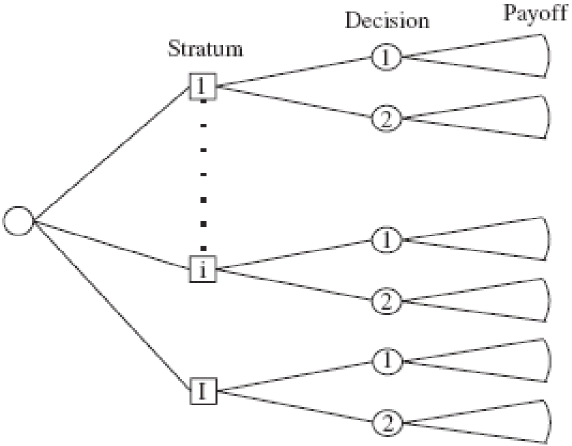
Why are textbook examples of decision analysis so lame?

- Your nephew is renting an apartment, balancing issues of cost, size, convenience, ...
- Widgets cost $2 to make and sell for $3. Here's the distribution of the market for widgets, ..., how many should you make?
- Vague business example
- Specific business example—what kind of power plant to build—pure GIGO
- Vague military example

Why are textbook examples of decision analysis so lame?

- Your nephew is renting an apartment, balancing issues of cost, size, convenience, . . .
- Widgets cost $2 to make and sell for $3. Here's the distribution of the market for widgets, . . . , how many should you make?
- Vague business example
- Specific business example—what kind of power plant to build—pure GIGO
- Vague military example

Technical challenges in evaluating decision trees



Evaluating nested decision trees

- Alternation of *decision nodes* and *uncertainty nodes*
- Evaluating a tree of decision nodes (e.g., a traffic route): maximize at each step
- Evaluating a tree of uncertainty nodes (e.g., a casino game): simulate random draw at each step
- Evaluating an alternating tree: difficult!

Evaluating nested decision trees

- Alternation of *decision nodes* and *uncertainty nodes*
- Evaluating a tree of decision nodes (e.g., a traffic route): maximize at each step
- Evaluating a tree of uncertainty nodes (e.g., a casino game): simulate random draw at each step
- Evaluating an alternating tree: difficult!

Evaluating nested decision trees

- Alternation of *decision nodes* and *uncertainty nodes*
- Evaluating a tree of decision nodes (e.g., a traffic route): maximize at each step
- Evaluating a tree of uncertainty nodes (e.g., a casino game): simulate random draw at each step
- Evaluating an alternating tree: difficult!

Evaluating nested decision trees

- Alternation of *decision nodes* and *uncertainty nodes*
- Evaluating a tree of decision nodes (e.g., a traffic route): maximize at each step
- Evaluating a tree of uncertainty nodes (e.g., a casino game): simulate random draw at each step
- Evaluating an alternating tree: difficult!

Evaluating nested decision trees

- Alternation of *decision nodes* and *uncertainty nodes*
- Evaluating a tree of decision nodes (e.g., a traffic route): maximize at each step
- Evaluating a tree of uncertainty nodes (e.g., a casino game): simulate random draw at each step
- Evaluating an alternating tree: difficult!

Institutional decision analysis

- Comparative decisions
- Understanding decision makers' priorities
- Relative recommendations

Institutional decision analysis

- Comparative decisions
- Understanding decision makers' priorities
- Relative recommendations

Institutional decision analysis

- Comparative decisions
- Understanding decision makers' priorities
- Relative recommendations

Institutional decision analysis

- Comparative decisions
- Understanding decision makers' priorities
- Relative recommendations

Decentralized decision making

- What is the role of the goverment/NGO?
- Coordinating data collection
- Centralized data analysis
- Providing individualized recommendations
- Hierarchical modeling for dispersed decision making

Decentralized decision making

- What is the role of the goverment/NGO?
- Coordinating data collection
- Centralized data analysis
- Providing individualized recommendations
- Hierarchical modeling for dispersed decision making

Decentralized decision making

- What is the role of the goverment/NGO?
- Coordinating data collection
- Centralized data analysis
- Providing individualized recommendations
- Hierarchical modeling for dispersed decision making

Decentralized decision making

- What is the role of the goverment/NGO?
- Coordinating data collection
- Centralized data analysis
- Providing individualized recommendations
- Hierarchical modeling for dispersed decision making

Decentralized decision making

- What is the role of the goverment/NGO?
- Coordinating data collection
- Centralized data analysis
- Providing individualized recommendations
- Hierarchical modeling for dispersed decision making

Decentralized decision making

- What is the role of the goverment/NGO?
- Coordinating data collection
- Centralized data analysis
- Providing individualized recommendations
- Hierarchical modeling for dispersed decision making

Advances in logistic regression

- Bayesian inference: the best fit to data does not give the best prediction for future data
- Conservatism in statistical inference
- Predictive model checking

Advances in logistic regression

- Bayesian inference: the best fit to data does not give the best prediction for future data
- Conservatism in statistical inference
- Predictive model checking

Advances in logistic regression

- Bayesian inference: the best fit to data does not give the best prediction for future data
- Conservatism in statistical inference
- Predictive model checking

Advances in logistic regression

- Bayesian inference: the best fit to data does not give the best prediction for future data
- Conservatism in statistical inference
- Predictive model checking

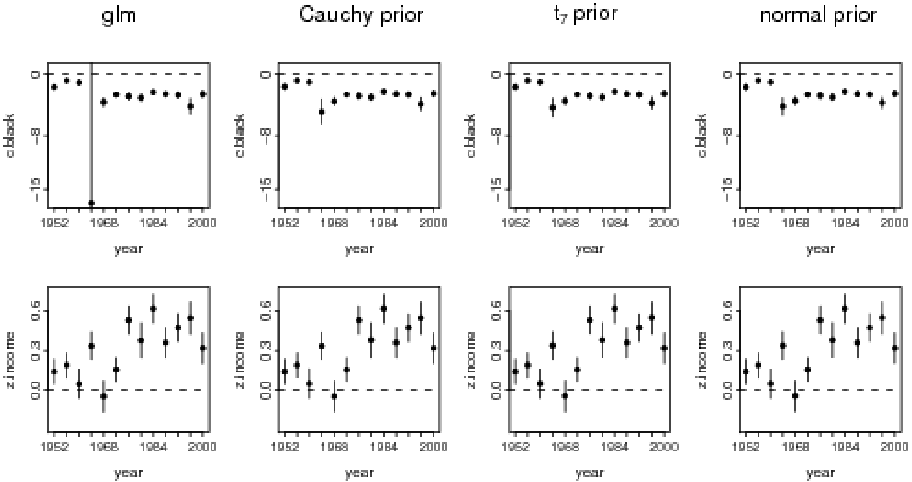
Separation in logistic regression

```
glm (vote ~ female + black + income, family=binomial(link="logit"))

1960                                    1968
             coef.est  coef.se                       coef.est  coef.se
(Intercept) -0.14       0.23            (Intercept)   0.47      0.24
female       0.24       0.14            female       -0.01      0.15
black       -1.03       0.36            black        -3.64      0.59
income       0.03       0.06            income       -0.03      0.07

1964                                    1972
             coef.est  coef.se                       coef.est  coef.se
(Intercept)  -1.15       0.22           (Intercept)   0.67      0.18
female       -0.09       0.14           female       -0.25      0.12
black       -16.83     420.40           black        -2.63      0.27
income        0.19       0.06           income        0.09      0.05
```

Regularization in action!



Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coefs are almost always between -5 and 5:
  - 5 on the logit scale takes you from 0.01 to 0.50 or from 0.50 to 0.99
    Smoking and lung cancer
- Independent Cauchy prior dists with center 0 and scale 2.5
- Rescale each predictor to have mean 0 and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of `glm`

Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coefs are almost always between −5 and 5:
  - 5 on the logit scale takes you from 0.01 to 0.50 or from 0.50 to 0.99
  - Smoking and lung cancer
- Independent Cauchy prior dists with center 0 and scale 2.5
- Rescale each predictor to have mean 0 and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of glm

Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coefs are almost always between -5 and 5:
  - 5 on the logit scale takes you from 0.01 to 0.50 or from 0.50 to 0.99
  - Smoking and lung cancer
- Independent Cauchy prior dists with center 0 and scale 2.5
- Rescale each predictor to have mean 0 and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of glm

Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coefs are almost always between -5 and 5:
  - 5 on the logit scale takes you from 0.01 to 0.50 or from 0.50 to 0.99
  - Smoking and lung cancer
- Independent Cauchy prior dists with center 0 and scale 2.5
- Rescale each predictor to have mean 0 and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of glm

Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coefs are almost always between -5 and 5:
  - 5 on the logit scale takes you from 0.01 to 0.50 or from 0.50 to 0.99
  - Smoking and lung cancer
- Independent Cauchy prior dists with center 0 and scale 2.5
- Rescale each predictor to have mean 0 and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of `glm`

Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coefs are almost always between -5 and 5:
  - 5 on the logit scale takes you from 0.01 to 0.50 or from 0.50 to 0.99
  - Smoking and lung cancer
- Independent Cauchy prior dists with center 0 and scale 2.5
- Rescale each predictor to have mean 0 and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of `glm`

Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coefs are almost always between -5 and 5:
  - 5 on the logit scale takes you from 0.01 to 0.50 or from 0.50 to 0.99
  - Smoking and lung cancer
- Independent Cauchy prior dists with center 0 and scale 2.5
- Rescale each predictor to have mean 0 and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of `glm`

Weakly informative priors for logistic regression coefficients

- Separation in logistic regression
- Some prior info: logistic regression coefs are almost always between -5 and 5:
  - 5 on the logit scale takes you from 0.01 to 0.50 or from 0.50 to 0.99
  - Smoking and lung cancer
- Independent Cauchy prior dists with center 0 and scale 2.5
- Rescale each predictor to have mean 0 and sd $\frac{1}{2}$
- Fast implementation using EM; easy adaptation of `glm`

Evaluation using a corpus of datasets

- Compare classical glm to Bayesian estimates using various prior distributions
- Evaluate using cross-validation and average predictive error
- The optimal prior distribution for β's is (approx) Cauchy $(0, 1)$
- Our Cauchy $(0, 2.5)$ prior distribution is weakly informative!

Evaluation using a corpus of datasets

- Compare classical glm to Bayesian estimates using various prior distributions
- Evaluate using cross-validation and average predictive error
- The optimal prior distribution for β's is (approx) Cauchy $(0, 1)$
- Our Cauchy $(0, 2.5)$ prior distribution is weakly informative!

Evaluation using a corpus of datasets

- Compare classical glm to Bayesian estimates using various prior distributions
- Evaluate using cross-validation and average predictive error
- The optimal prior distribution for β's is (approx) Cauchy $(0, 1)$
- Our Cauchy $(0, 2.5)$ prior distribution is weakly informative!

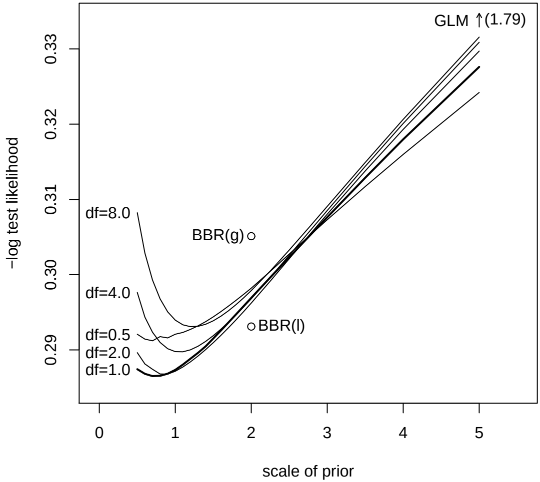
Evaluation using a corpus of datasets

- Compare classical glm to Bayesian estimates using various prior distributions
- Evaluate using cross-validation and average predictive error
- The optimal prior distribution for β's is (approx) Cauchy $(0, 1)$
- Our Cauchy $(0, 2.5)$ prior distribution is weakly informative!

Evaluation using a corpus of datasets

- Compare classical glm to Bayesian estimates using various prior distributions
- Evaluate using cross-validation and average predictive error
- The optimal prior distribution for β's is (approx) Cauchy $(0, 1)$
- Our Cauchy $(0, 2.5)$ prior distribution is weakly informative!

Expected predictive loss, avg over a corpus of datasets



Conservatism of Bayesian inference

- Consider the logistic regression example
- Problems with maximum likelihood when data show separation:
  - Coefficient estimate of $-\infty$
  - Estimated predictive probability of 0 for new cases
- Is this conservative?
- Not if evaluated on new data
- What is statistical conservatism?

Conservatism of Bayesian inference

- Consider the logistic regression example
- Problems with maximum likelihood when data show separation:
  - Coefficient estimate of $-\infty$
  - Estimated predictive probability of 0 for new cases
- Is this conservative?
- Not if evaluated on new data
- What is statistical conservatism?

Conservatism of Bayesian inference

- Consider the logistic regression example
- Problems with maximum likelihood when data show separation:
  - Coefficient estimate of $-\infty$
  - Estimated predictive probability of 0 for new cases
- Is this conservative?
- Not if evaluated on new data
- What is statistical conservatism?

Conservatism of Bayesian inference

- Consider the logistic regression example
- Problems with maximum likelihood when data show separation:
  - Coefficient estimate of $-\infty$
  - Estimated predictive probability of 0 for new cases
- Is this conservative?
- Not if evaluated on new data
- What is statistical conservatism?

Conservatism of Bayesian inference

- Consider the logistic regression example
- Problems with maximum likelihood when data show separation:
  - Coefficient estimate of $-\infty$
  - Estimated predictive probability of 0 for new cases
- Is this conservative?
- Not if evaluated on new data
- What is statistical conservatism?

Conservatism of Bayesian inference

- Consider the logistic regression example
- Problems with maximum likelihood when data show separation:
  - Coefficient estimate of $-\infty$
  - Estimated predictive probability of 0 for new cases
- Is this conservative?
- Not if evaluated on new data
- What is statistical conservatism?

Conservatism of Bayesian inference

- Consider the logistic regression example
- Problems with maximum likelihood when data show separation:
  - Coefficient estimate of $-\infty$
  - Estimated predictive probability of 0 for new cases
- Is this conservative?
- Not if evaluated on new data
- What is statistical conservatism?

Conservatism of Bayesian inference

- Consider the logistic regression example
- Problems with maximum likelihood when data show separation:
  - Coefficient estimate of $-\infty$
  - Estimated predictive probability of 0 for new cases
- Is this conservative?
- Not if evaluated on new data
- What is statistical conservatism?

Take-home points

- Classical tools for statistical analysis and decision making are being made more realistic
- Recognize and surmount the garbage-in, garbage-out nature of decision analysis and statistical modeling
- Thanks also to Lex van Geen, Matilde Trevisani, Jie Shen, Hao Lu, Erwann Rogard, and Aleks Jakulin

Take-home points

- Classical tools for statistical analysis and decision making are being made more realistic
- Recognize and surmount the garbage-in, garbage-out nature of decision analysis and statistical modeling
- Thanks also to Lex van Geen, Matilde Trevisani, Jie Shen, Hao Lu, Erwann Rogard, and Aleks Jakulin

Take-home points

- Classical tools for statistical analysis and decision making are being made more realistic
- Recognize and surmount the garbage-in, garbage-out nature of decision analysis and statistical modeling
- Thanks also to Lex van Geen, Matilde Trevisani, Jie Shen, Hao Lu, Erwann Rogard, and Aleks Jakulin

Take-home points

- Classical tools for statistical analysis and decision making are being made more realistic
- Recognize and surmount the garbage-in, garbage-out nature of decision analysis and statistical modeling
- Thanks also to Lex van Geen, Matilde Trevisani, Jie Shen, Hao Lu, Erwann Rogard, and Aleks Jakulin