Progress in Risk Science and Causality

Tony Cox, tcoxdenver@aol.com

AAPCA

March 27, 2017

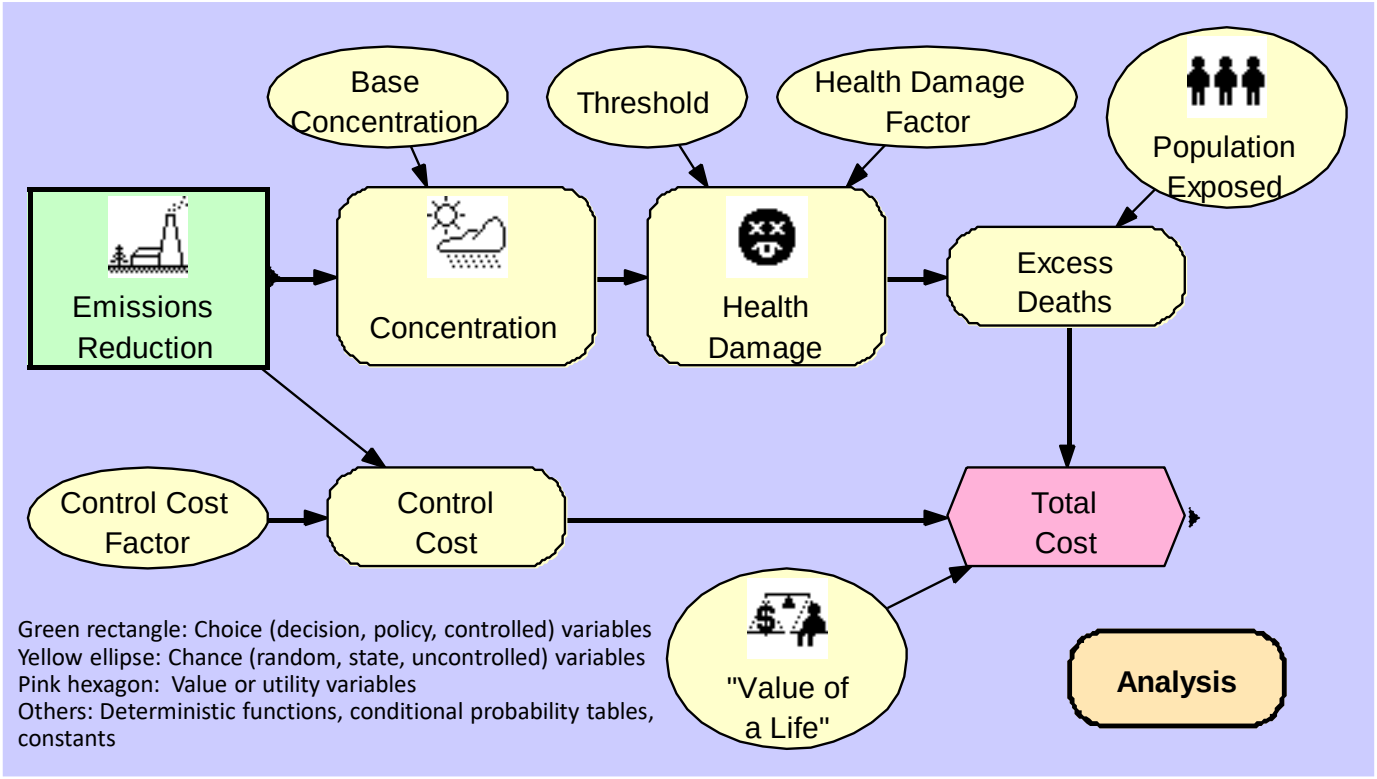
Vision for causal analytics

- Represent understanding of how the world works by an explicit *causal model.*
  - Learn, validate, and document models with data
- Use causal model to quantify how probabilities of *consequences* change as *decision variables* or *policies* are changed
- Given preferences, *solve for best policy*
  - Choose policy variables to cause maximum net benefit (or expected social utility)
  - Perform sensitivity analyses, value-of-information (VOI) analyses; optimize timing of interventions
  - Evaluate results, adaptively learn and improve policies

Representing understanding via a causal graph (DAG)

Example: *Analytica* Influence Diagram (ID) causal model

Total Cost to society = *Control Cost* of *Emissions Reduction* + *"Value of a Life"*Excess Deaths*



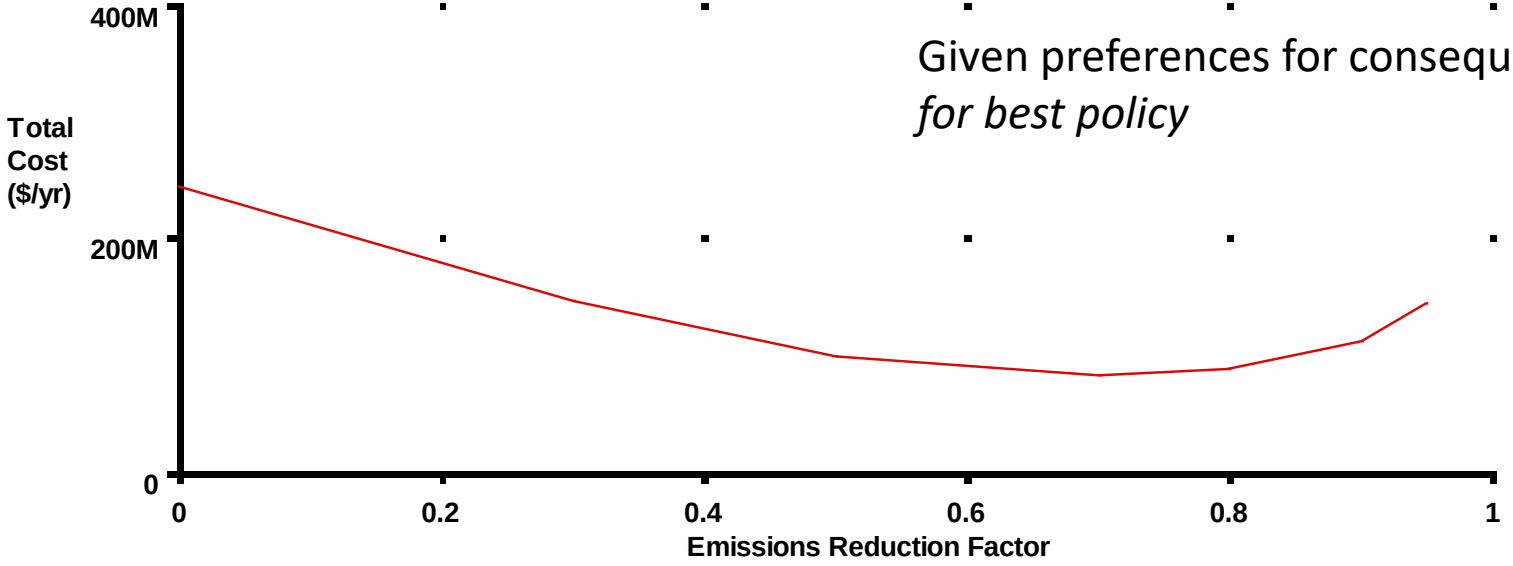
Modular structure

- Each variable is *conditionally independent* of its more remote ancestors, given the values of its parents in the "DAG" (directed acyclic graph)
- *Dependencies* are quantified by *conditional probability tables* (CPTs) or trees

Using the causal model to solve for the best policy, via probabilistic simulation of outcomes

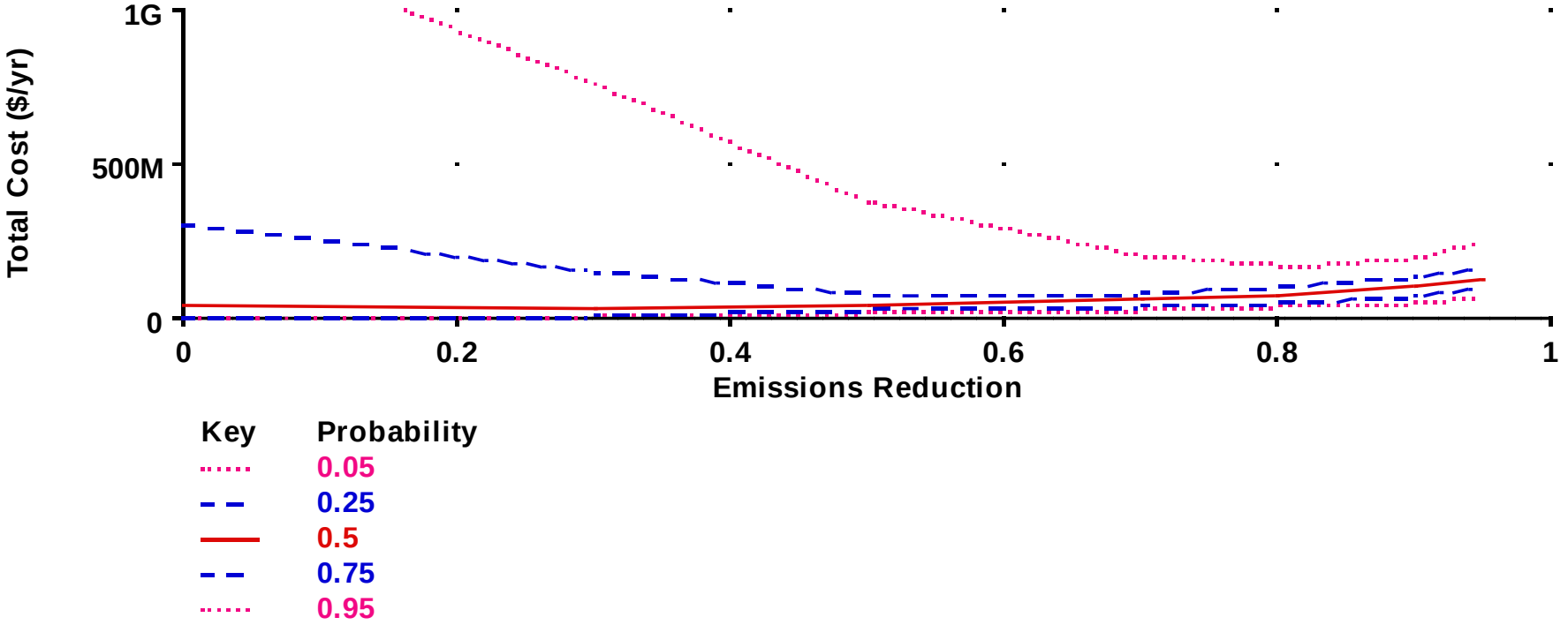
Use causal model to quantify how probabilities of *consequences* change as *decision variables* or *policies* are changed

Given preferences for consequences, *solve for best policy*



Simulation-based *partial dependence plot*: Decision variable is varied over a range of alternative (counterfactual) values. Other variables are drawn from distributions, for each value of the decision variable.

Analytica's probabilistic simulation-based uncertainty analysis



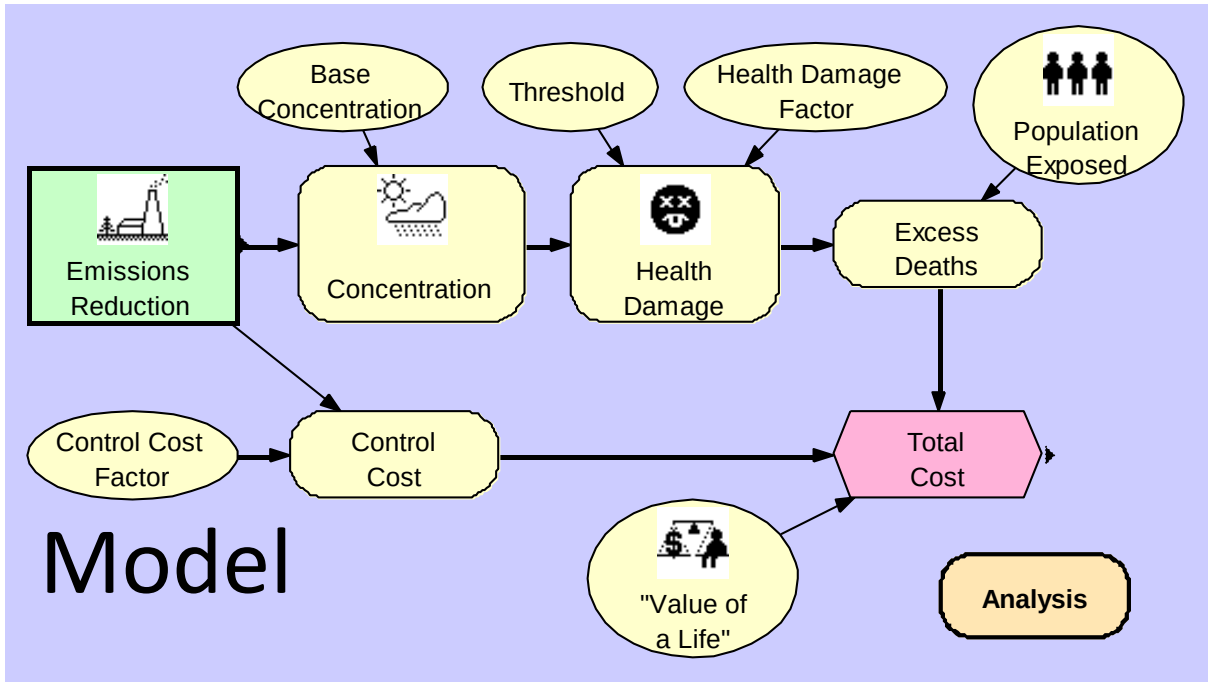
If correct causal model is known, then it can be used to support risk management policy decisions.

Causal analytics: How to learn and validate causal models *from data*?

year	month	day	AllCause75	PM2.5	tmin	tmax	MAXRH
2007	1	1	151	38.4	36	72	68.8
2007	1	2	158	17.4	36	75	48.9
2007	1	3	139	19.9	44	75	61.3
2007	1	4	164	64.6	37	68	87.9
2007	1	5	136	6.1	40	61	47.5
2007	1	6	152	18.8	39	69	39
2007	1	7	171	19.1	41	76	40.9
2007	1	8	164	13.8	41	83	33.7
2007	1	9	188	14.6	41	84	37.5
2007	1	10	169	39.6	41	78	63.2
2007	1	11	160	19.2	37	66	85.9
2007	1	12	160	22.3	31	56	67.2
2007	1	13	166	11.7	27	55	40.4

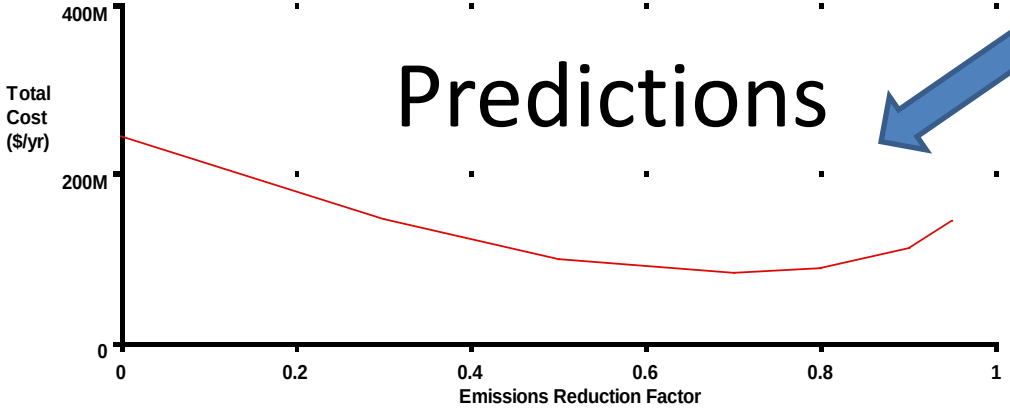
Data

?



Predictions

Monte Carlo simulation



What we want from causal analysis

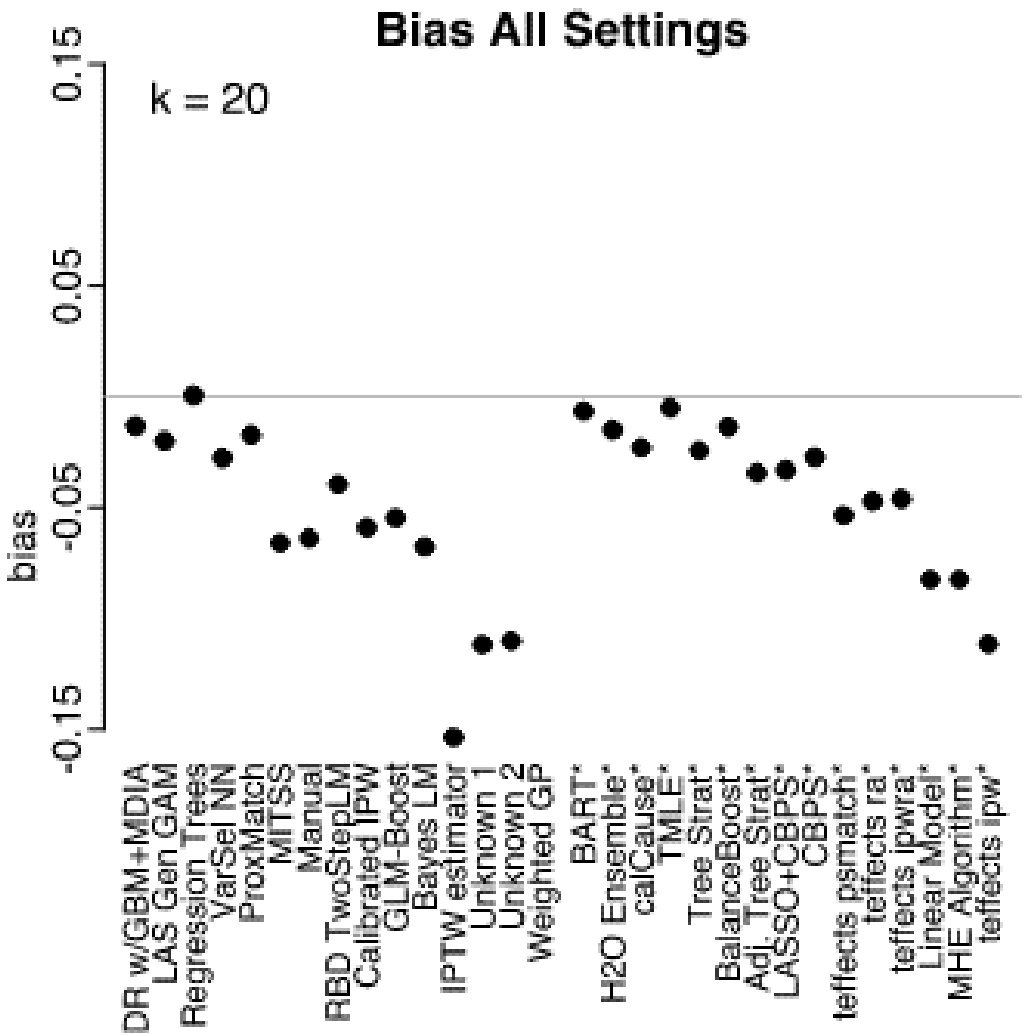
- ***Predict/evaluate*** consequences of policies
  - How will (or how did) *actions* change frequencies of *outcomes* in a population?
    - How does reducing exposure change risks?
    - Manipulative causality
    - *Not* associational, attributive, counterfactual/ potential outcomes, predictive, or mechanistic causality
  - Most existing air pollution health effects studies and regulatory benefits assessments do not address predictive or manipulative causality (Schwartz et al., 2016)
    - Decades of associational, attributive, counterfactual/ potential outcomes, and mechanistic studies

What we can get from data: Predictive causality

- Do changes in exposures *help to predict* changes in health effects?
  - Can be answered objectively from observational data without assumptions using machine-learning algorithms
    - Data from valid quasi-experimental designs
    - Analytics via information-based causal discovery algorithms
    - Automated, well-vetted and validated software in R and Python
  - *Predictive* causation is necessary (but not sufficient) for what we want: *manipulative* causation
    - Useful screening tool , errs on the side of false positives
    - Counterexample: Nicotine-stained fingers are a predictive cause of lung cancer, but not a manipulative or mechanistic cause

How to do it: Information-based causal discovery algorithms dominate competitive evaluations

(Hill, 2016, http://jenniferhill7.wixsite.com/acic-2016/competition)

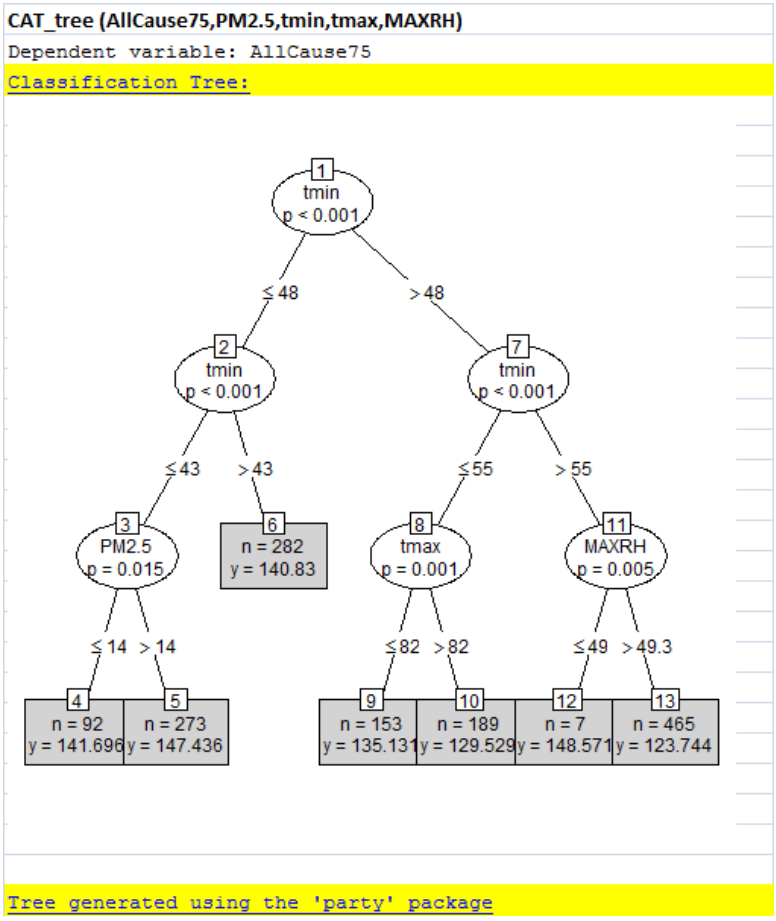


Empirically, tree-based model ensembles do best in causal discovery algorithm competitions

Propensity scoring (potential outcome models) do worst (about 20 times greater prediction error and bias)

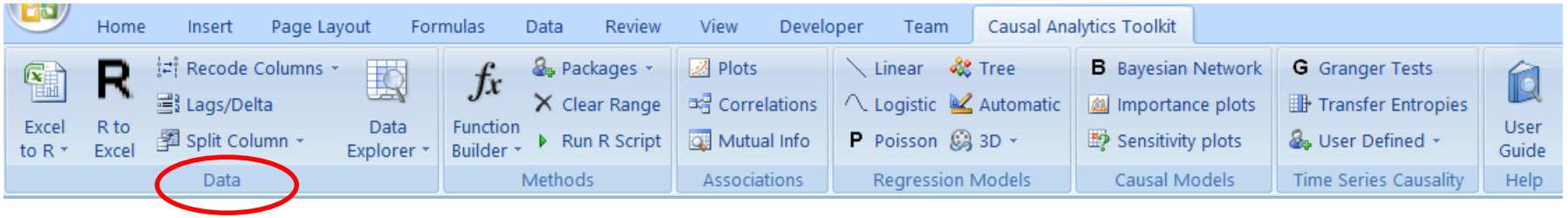
Principles of most successful causal effect estimation algorithms

- **Information** principle: Causes are informative about (help to predict) their effects
  - So, exploit predictive analytics algorithms!
    - Use DAGs, trees, Random Forests , etc. to find informative variables
  - Propagation principle: Changes in causes help to predict and explain changes in their effects
  - Information flows from causes to their effects over time
- Use **non-parametric** effects estimates
  - CART trees estimate conditional probabilities directly from data, no parametric model
    - Avoids errors from modeling biases, specification errors, uncertain assumptions
  - Allow nonlinearities and interactions
- Average over **ensembles** of hundreds of non-parametric estimates/predictions

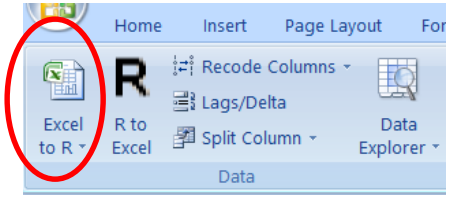


Making it easy:
Causal Analytics Toolkit (CAT)

http://cox-associates.com/downloads/



CAT applies R and Python analytics to data in Excel



- Load data in Excel, click *Excel to R* to send it to R
  - Los Angeles air basin
  - 1461 days, 2007-2010 (Lopiano et al., 2015, thanks to Stan Young for data)
  - PM2.5 data from CARB
  - Elderly mortality ("AllCause75") from CA Department of Health
  - Daily min and max temps & max relative humidity from ORNL and EPA

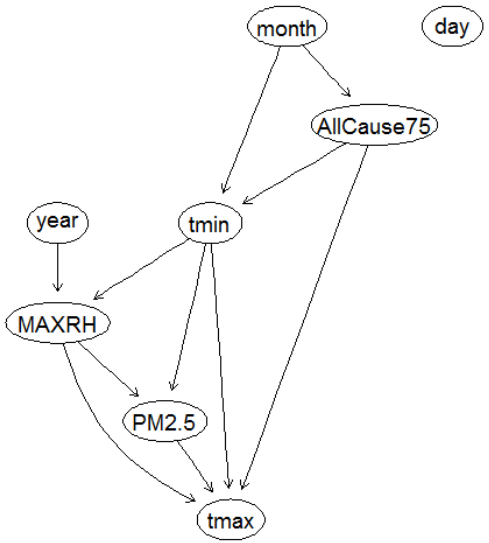
year	month	day	AllCause75	PM2.5	tmin	tmax	MAXRH
2007	1	1	151	38.4	36	72	68.8
2007	1	2	158	17.4	36	75	48.9
2007	1	3	139	19.9	44	75	61.3
2007	1	4	164	64.6	37	68	87.9
2007	1	5	136	6.1	40	61	47.5
2007	1	6	152	18.8	39	69	39
2007	1	7	160	19.1	41	76	40.9
2007	1	8	148	13.8	41	83	33.7
2007	1	9	188	14.6	41	84	37.5
2007	1	10	169	39.6	41	78	63.2
2007	1	11	160	19.2	37	66	85.9
2007	1	12	160	22.3	31	56	67.2
2007	1	13	166	11.7	27	55	40.4

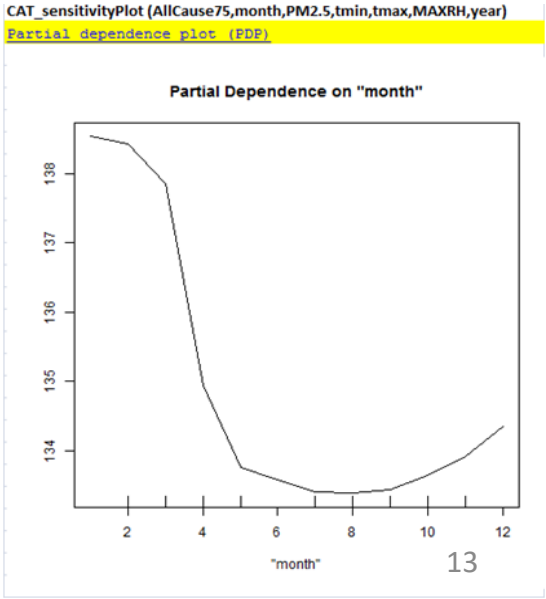
- *Risk question:* Does PM2.5 exposure increase elderly mortality risk? If so, how much?

Basic ideas of information-based Causal Analytics

- Use a (DAG) network to show which variables provide direct information about each other
  - Arrows between variables show they are *informative about* each other
    - Learn network structure directly from data
    - Scoring algorithms, constraint algorithms, hybrids
  - Carefully check conclusions
    - In non-parametric analyses we trust!
    - Do power analyses using simulation
  - Interpret neighbors in network as *potential direct causes* (satisfying necessary condition)

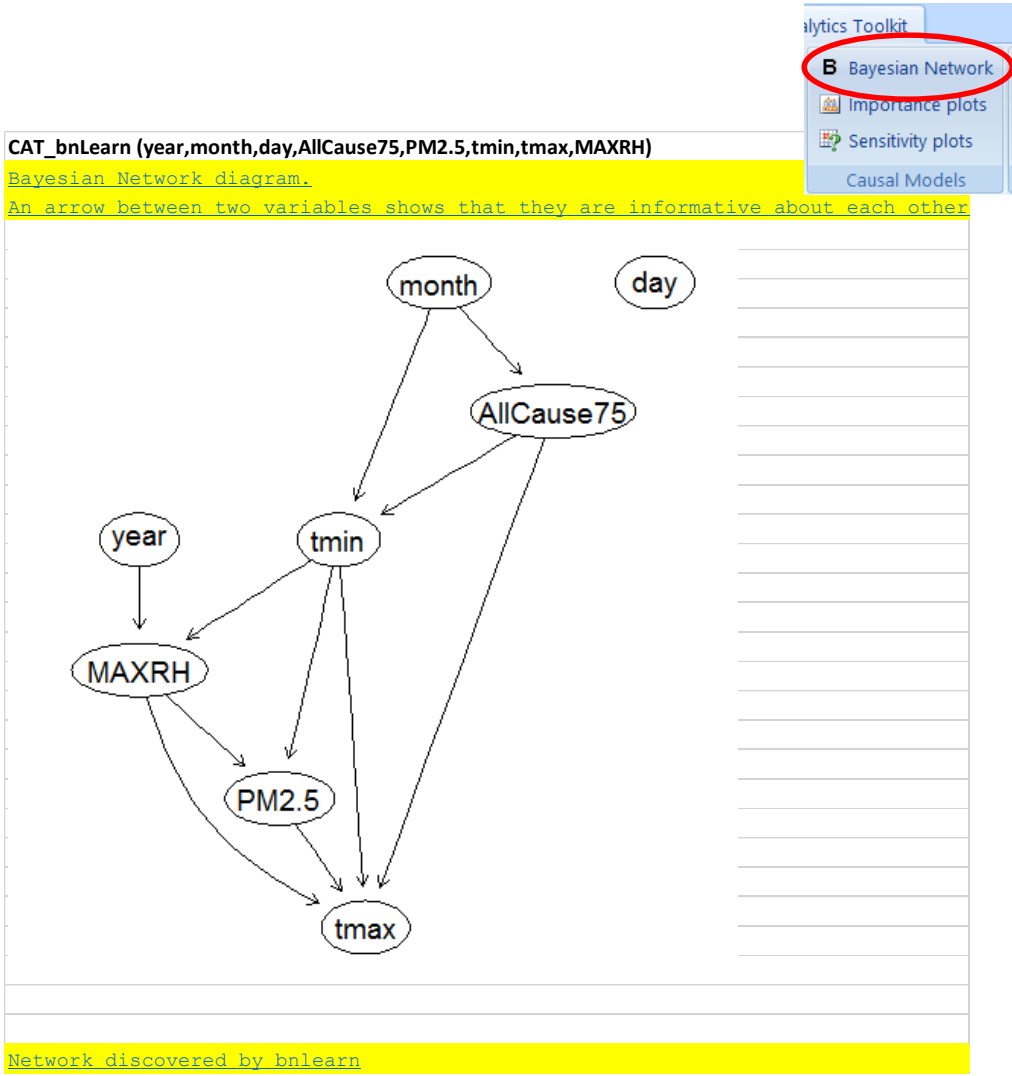
- Use partial dependence plots learned from data (based on averaging over many trees) to quantify relation between inputs and dependent variables.





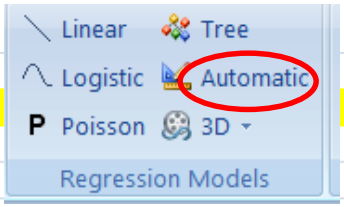
Run BN structure discovery algorithms

- Click **B** *Bayesian Network* to generate DAG structure.
  - Only variables connected to response variable by an arrow are identified as potential direct causes
  - Multiple pathways between two variables reveal potential direct and indirect effects
  - *Example:* Direct and indirect paths between tmax and AllCause75.



By contrast, regression estimates *total* associations, given an assumed model

- Click on *Automatic* under Regression Models
- CAT selects and runs appropriate regression models, reports results
  - Quasi-Poisson regression model shows *significant positive total C-R association between PM2.5 and elderly mortality* (AllCause75)
  - Significant *negative* total association between temperature and elderly mortality



Dependent variable: AllCause75

Quasi-Poisson regression model

Estimated Coefficients

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.684524	4.995490	0.74	0.4609	
PM2.5	0.000745	0.000254	2.93	0.0034	**
tmin	-0.003820	0.000626	-6.10	1.4e-09	***
tmax	-0.001776	0.000447	-3.98	7.3e-05	***
MAXRH	-0.000961	0.000235	-4.10	4.4e-05	***
year	0.000833	0.002489	0.33	0.7379	
month	-0.009686	0.000809	-11.98	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

95% Confidence Intervals

	2.5 %	97.5 %
(Intercept)	-6.106729	13.475263
PM2.5	0.000247	0.001241
tmin	-0.005048	-0.002593
tmax	-0.002651	-0.000901
MAXRH	-0.001421	-0.000501
year	-0.004044	0.005710
month	-0.011271	-0.008102

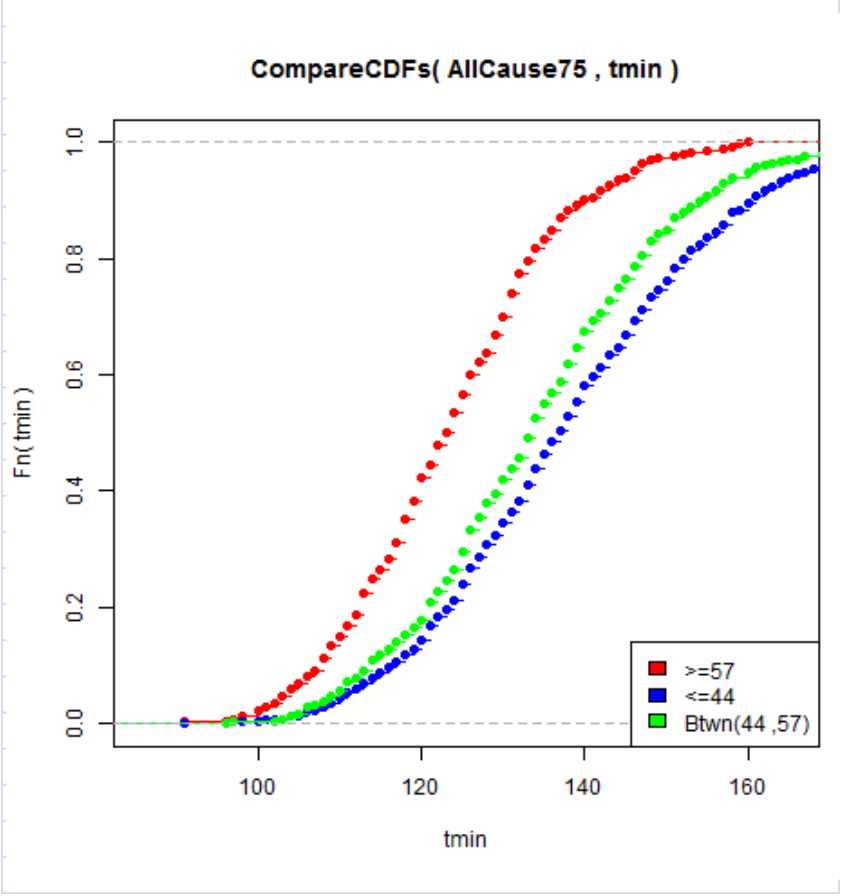
Confirm or refute/refine BN structure with additional non-parametric tests

- Conditioning on very different values of a direct cause should cause the distribution of the response variable to change
- If the response variable does not change, then any association between them may be due to indirect pathways (e.g., confounding)

CAT_compareCDFs(AllCause75,tmin)

Cumulative distribution function

tmin lower quartile(0.25 of values) = 44 , upper quartile(0.75 of values) = 57



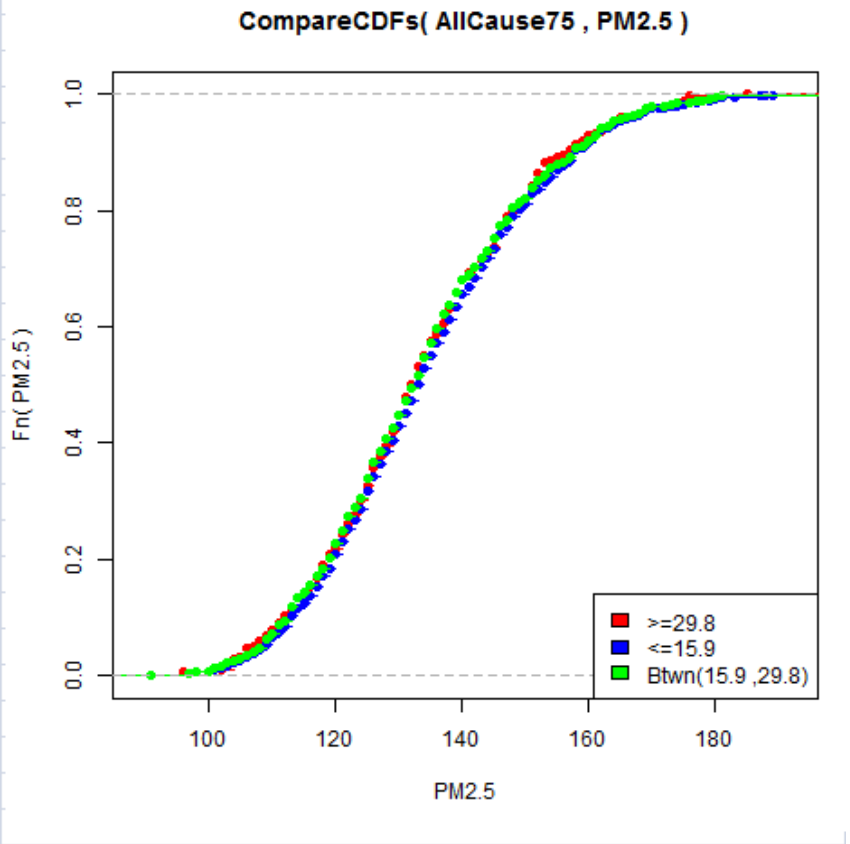
Confirm or refute/refine BN structure with additional non-parametric tests

- Conditioning on very different values of a direct cause should cause the distribution of the response variable to change
- If the response variable does not change, then any association between them may be due to indirect pathways (e.g., confounding)

CAT_compareCDFs(AllCause75,PM2.5)

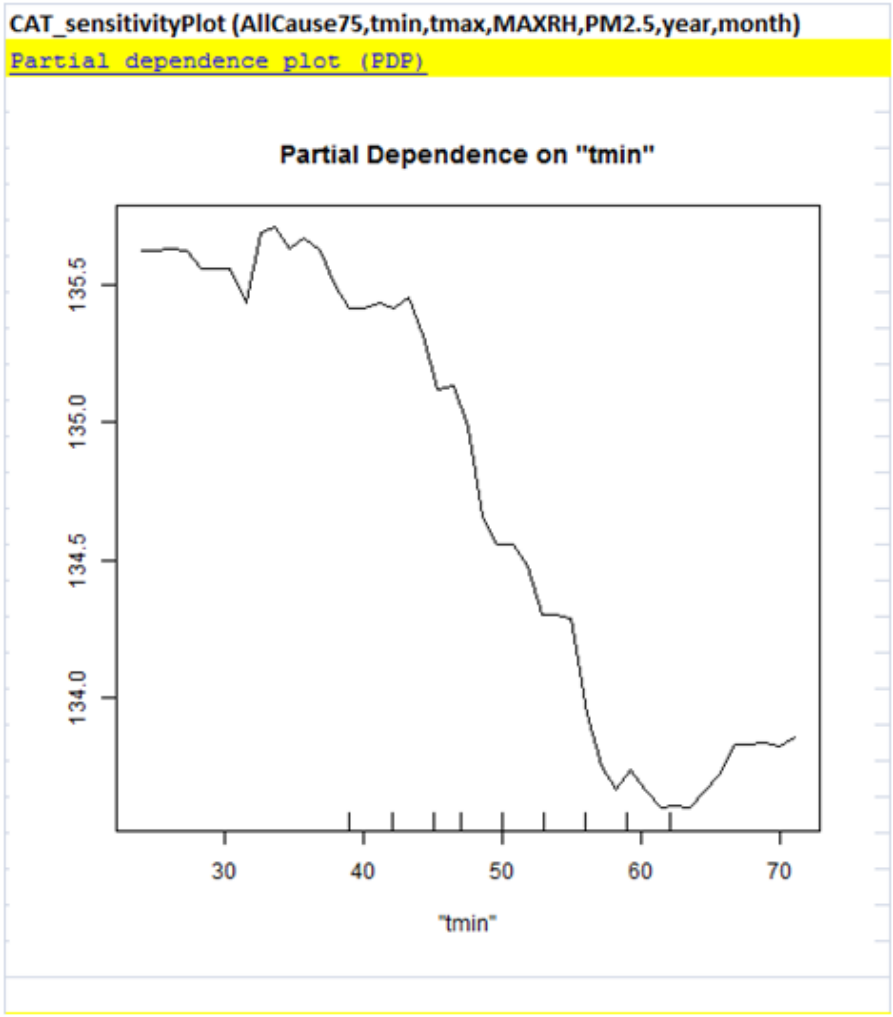
Cumulative distribution function

PM2.5 lower quartile(0.25 of values) = 15.9 , upper quartile(0.75 of values) = 29.8



Quantify direct causal relations

- *Procedure:* To quantify direct (potentially causal) relations after controlling for other variables, use *partial dependence plot* for response R vs. (potential) cause C.
  - RandomForest algorithm averages multiple independent conditional probability predictions of outcome Y for each value of x
  - *Rationale:* DAG structure shows that the relation *might* be causal (X helps to predict Y). Partial dependence estimates size of potential effect.
  - Data-based simulation of conditional expected values generates curve



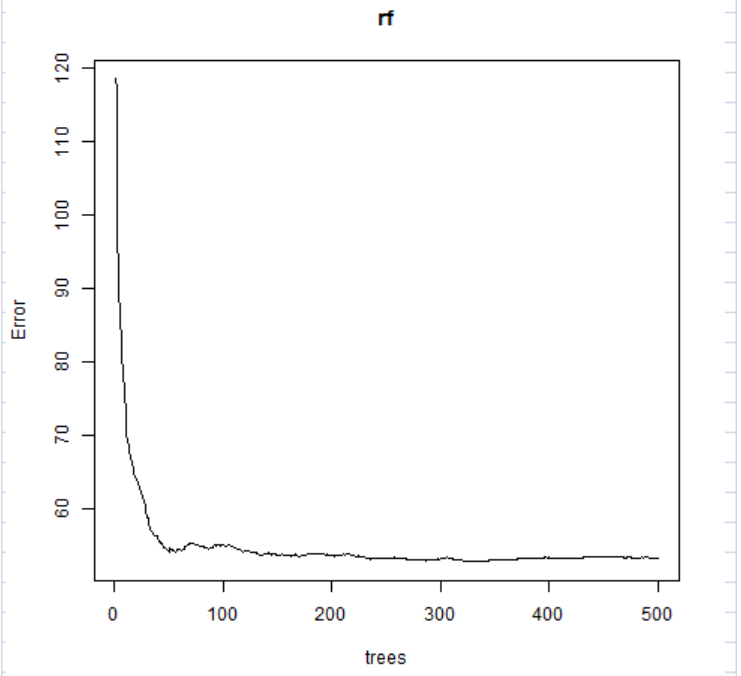
Validate quantified C-R relations in hold-out sample

- CAT currently quantifies uncertainty using bootstrap and cross-validation approaches for Random Forest ensembles
- Averaging over many trees reduces MSE (mean squared prediction error)

```
CAT_importance (tmax,MAXRH,year,month,AllCause75,PM2.5)
Dependent variable: tmax
randomForest summary
No. of variables tried at each split: 1

          Mean of squared residuals: 53.31128
                    % Var explained: 69.36

randomForest accuracy vs. number of trees
```



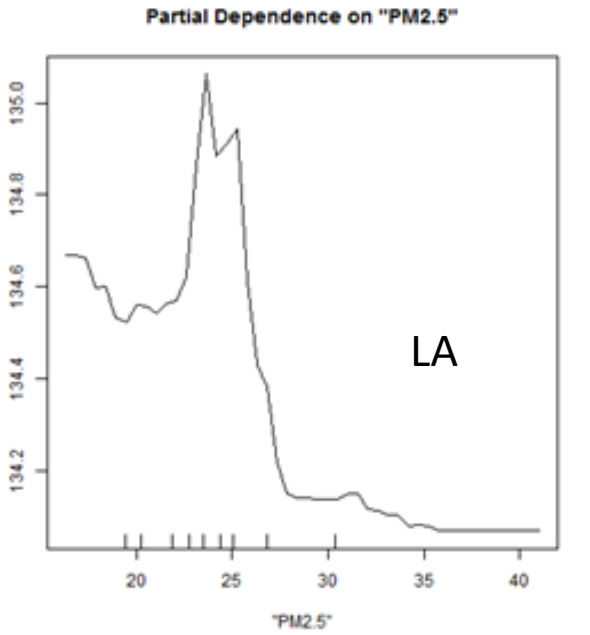
Example applications: Boston & LA areas

CAT_sensitivityPlot (mortality_75,PM2.5,Tmin,Tmax,MAXRH,year,month

Dependent variable: mortality_75

Partial dependence plot (PDP)

Same plot with different ranges of y-axis

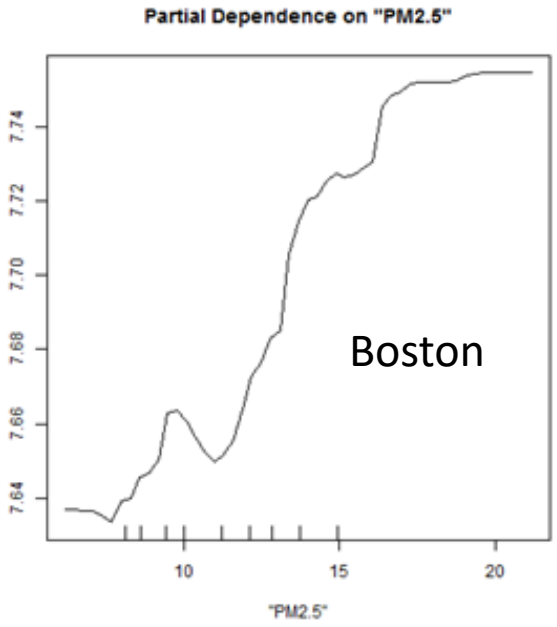


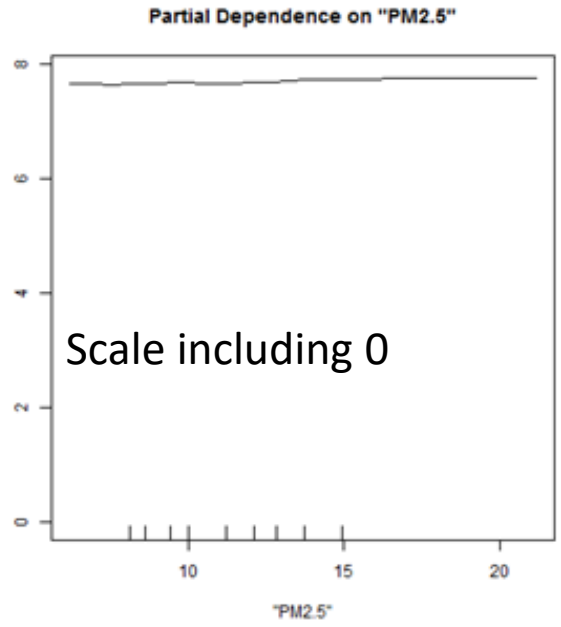
CAT_sensitivityPlot (mortality75,PM2.5,Tmax,Tmin,Dewpoint,year,month,time)

Dependent variable: mortality75

Partial dependence plot (PDP)

Same plot with different ranges of y-axis





- PM2.5 has positive regression coefficient as predictor of AllCause75 in both data sets
- PM2.5 is not a significant predictive causal predictor in either data set
- C-R function learned from Boston does not apply in LA (and *vice versa*)

Wrap-up on CAT

- Modern software makes it easy to apply information-based causal analytics to epidemiological data
- Entire analysis process can be automated
  - Click on “Analyze” in CAT
  - Minimizes roles of modeling choices, p-hacking, confirmation bias, etc.
- Limited but useful outputs: Possible causal relationships detected and quantified directly from data
  - Predictive causality, not necessarily manipulative

Recent advances in causal analytics for risk analysis

- Clarifying what we want
  - Manipulative causation
- Identifying what we can get
  - Predictive causation
- Understanding how to get it
  - Information-based causal discovery algorithms
- Understanding how not to get it
  - Associations, unverified assumptions, judgments
- Comparative evaluation and validation of causal discovery algorithms

Literature has usually not addressed predictive (or manipulative) causation

- How *not* to assess predictive causation
  - Statistical associations (almost all existing literature)
  - Judgments, weight-of-evidence (IARC)
  - Untested modeling assumptions
    - Potential outcome/counterfactual models, instrumental variables, regression discontinuity designs, etc.
    - Intervention studies without control groups

- These methods are technically outdated (non-competitive), but are still dominant in practice

What we want from causal analysis

- ***Predict/Evaluate*** effects of actions
- ***Objectivity:*** Answers determined by data
  - Answers should not depend on modeling choices
  - Answers should not depend on judgments (as at IARC)
  - Discover (don't assume) how actions affect outcomes
  - "Putting the science back into risk science"
- ***Generalizability:*** Adjust answers across locations
- ***Uncertainty characterization:*** How sure can we be that policy changes cause desired effects?
  - Not "How likely is association to be causal?"
  - Value of information: What would reduce uncertainty?
- ***Performance:*** Validate how well methods work

What information-based causal analytics gives us

- ***Predict/Evaluate*** based on predictive causation
- ***Objectivity:*** Answers determined by data + predictive analytics machine learning (ML) algorithms
  - No untested assumptions
  - Non-parametric test: Is future of Y conditionally independent of history of X, given the history of Y? (Granger causation)
- ***Generalizability:*** Adjust answers across locations by applying causal CPT to different contexts, z
  - "Transport formulas" for generalizing causal findings
- ***Uncertainty characterization:*** Use non-parametric *model ensembles* (e.g., Random Forest algorithm from ML) to quantify cross-validated prediction error rates
- ***Performance validation:*** Cross-validation, competition
  - Extensive ML literature, competitive evaluations

Summary

- Modern software makes predictive and causal analytics easy to apply
- What we want: Manipulative causation
- What we can get: Predictive causation
- How: Information-based causal discovery algorithms
  - No longer needed: Associations, unverified assumptions, judgments
- Comparative evaluation and validation: Information-based causal discovery explains associations well in practice
- Can successfully identify and quantify possible causal relationships from data

Thanks!