Regression Models You Can See (and Interpret)

Wei-Yin Loh

Department of Statistics
University of Wisconsin — Madison

Outline

- Introduction
  - Bane of regression modeling
  - Boston housing data example
- GUIDE — an approach to visualizable regression models
  - Simple linear regression trees
    - ★ Resolving ambiguities in multiple linear regression
  - Two-predictor linear regression trees
    - ★ Application to outlier detection
- Empirical comparison of prediction accuracy
  - 27 algorithms and 52 datasets
  - Results
- Conclusion

The bane of regression modeling

- Numerous regression methods exist
- Many have good prediction accuracy
- Few yield interpretable models
- Very few (none?) are both accurate and interpretable
- Classical multiple linear regression model may be accurate — but it is often harder to interpret than might be expected

1970 Boston housing data

Var.	Definition	Var.	Definition
TOWN	township (92 values)	ID	census tract number
MEDV	median value in \$1000	AGE	% built before 1940
CRIM	per capita crime rate	DIS	dist. employ. centers
ZN	% land zoned for lots	RAD	access. to radial hwys
INDUS	% nonretail business	TAX	property tax rate/\$10K
CHAS	1 on river, 0 else	PT	pupil/teacher ratio
NOX	nitrogen oxide (p.p. 10^9)	B	(% black - $63)^2/10$
LSTAT	% lower-status pop.	RM	ave. number of rooms

Data: 506 observations (census tracts) in greater Boston

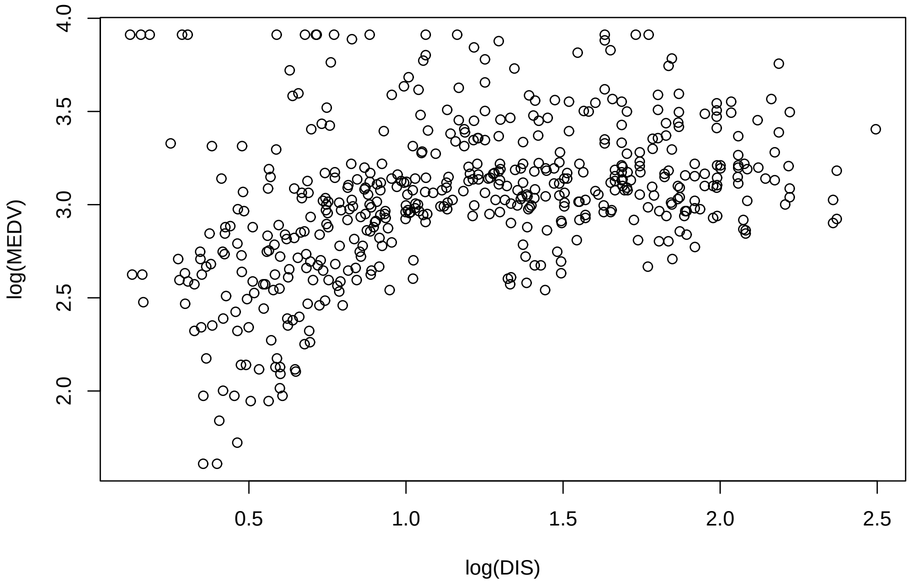
Goal: Examine the impact of air pollution on house price

Sources: Harrison & Rubinfeld (1978); Belsley, Kuh & Welsch (1980)

Harrison & Rubinfeld linear model for log(MEDV)

Variable	Coef	*t*-stat	Variable	Coef	*t*-stat
	4.6	30.0	AGE	7.1E-5	0.1
CRIM	-1.2E-2	-9.6	log(DIS)	-2.0E-1	-6.0
ZN	9.2E-5	0.2	log(RAD)	9.0E-2	4.7
TAX	-4.2E-4	-3.5	INDUS	1.8E-4	0.1
CHAS	9.2E-2	2.8	PT	-3.0E-2	-6.0
NOX^2	-6.4E-1	-5.7	B	3.6E-4	3.6
RM^2	6.3E-3	4.8	log(LSTAT)	-3.7E-1	-15.2

log(MEDV) vs. log(DIS)



Harrison & Rubinfeld linear model for log(MEDV)

X	β	t	ρ	X	β	t	ρ
	4.6	30.0		AGE	7.1E-5	0.1	-0.5
CRIM	-1.2E-2	-9.6	-0.5	log(DIS)*	-2.0E-1	-6.0	0.4
ZN	9.2E-5	0.2	0.4	log(RAD)*	9.0E-2	4.7	-0.4
TAX	-4.2E-4	-3.5	-0.6	INDUS	1.8E-4	0.1	-0.5
CHAS	9.2E-2	2.8	0.2	PT	-3.0E-2	-6.0	-0.5
NOX^2	-6.4E-1	-5.7	-0.5	B	3.6E-4	3.6	0.4
RM^2	6.3E-3	4.8	0.6	log(LSTAT)	-3.7E-1	-15.2	-0.8

β = coefficient, t = t-statistic, ρ = corr(X, Y)

Which sign is right?

- Coefficient from a multiple linear (ML) model is more trustworthy than that from a simple linear model because it adjusts for the effects of the other predictors

Which sign is right?

- Coefficient from a multiple linear (ML) model is more trustworthy than that from a simple linear model because it adjusts for the effects of the other predictors
- But the coefficient from the ML model depends on the form and number of the other predictors in the model

Which sign is right?

- Coefficient from a multiple linear (ML) model is more trustworthy than that from a simple linear model because it adjusts for the effects of the other predictors
- But the coefficient from the ML model depends on the form and number of the other predictors in the model
- If the ML model is wrong, the signs of its coefficients may be wrong too

How to avoid contradictory signs?

Overly simplistic solution

Choose one predictor variable and use simple linear regression

How to avoid contradictory signs?

Overly simplistic solution

Choose one predictor variable and use simple linear regression

Practical solution

- Partition the data until at most one or two predictors affect the response in each partition
- Fit a one- or two-predictor model to each partition
- Partitioning has the effect of conditioning on the other predictors

How to avoid contradictory signs?

Overly simplistic solution

Choose one predictor variable and use simple linear regression

Practical solution

- Partition the data until at most one or two predictors affect the response in each partition
- Fit a one- or two-predictor model to each partition
- Partitioning has the effect of conditioning on the other predictors

Why will two predictors not cause problems?

Because interpretation need not depend on the coefficients
— model and data can be visualized graphically

How to partition?

How to partition?

GUIDE — Loh (2002) *Statistica Sinica*, **12**, 361–386

Recursively, at each step:

1. Fit the best simple or two-predictor linear model to the data
2. Use residuals to choose the most nonlinear predictor
3. Partition the data with the chosen predictor variable

GUIDE split variable selection by chi-square analysis of residual patterns

1. Divide observations into two classes by signs of their residuals

GUIDE split variable selection by chi-square analysis of residual patterns

1. Divide observations into two classes by signs of their residuals
2. Curvature tests
   1. Discretize each continuous predictor and cross-classify against residual signs
   2. Find the p-value of each chi-square test

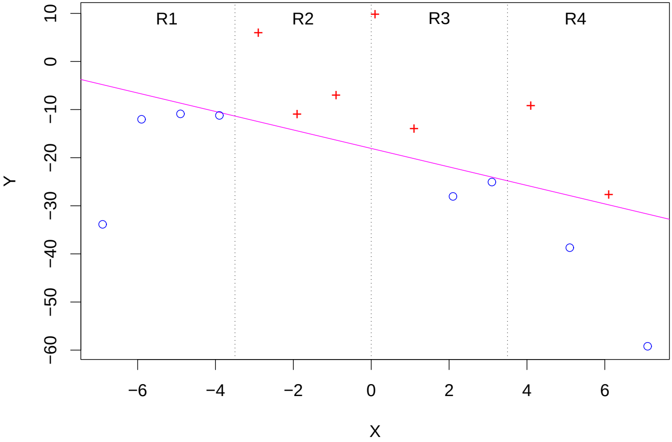
GUIDE split variable selection by chi-square analysis of residual patterns

1. Divide observations into two classes by signs of their residuals
2. Curvature tests
   1. Discretize each continuous predictor and cross-classify against residual signs
   2. Find the p-value of each chi-square test
3. Pairwise interaction tests
   1. Divide the 2D-space of each pair of predictor variables into groups
   2. Find the p-value of each chi-square test of residual signs vs. groups

GUIDE split variable selection by chi-square analysis of residual patterns

1. Divide observations into two classes by signs of their residuals
2. Curvature tests
    1. Discretize each continuous predictor and cross-classify against residual signs
    2. Find the p-value of each chi-square test
3. Pairwise interaction tests
    1. Divide the 2D-space of each pair of predictor variables into groups
    2. Find the p-value of each chi-square test of residual signs vs. groups
4. Split node with variable having the smallest p-value

Curvature test



Residual sign	R1	R2	R3	R4
Positive	0	3	2	2
Negative	4	0	2	2

When to stop partitioning?

- Stop only when data get too thin
- Partitioning generates a binary tree structure which in turn yields a nested sequence of piecewise linear models
- Select a piecewise model from the sequence by estimating the prediction error of each with cross-validation or an independent test sample

Advantages of GUIDE approach

Why use residuals to choose split variables?

- Unbiasedness in variable selection
- Savings in computation time
- Extensibility to robust, quantile, Poisson, relative risk, etc., regression

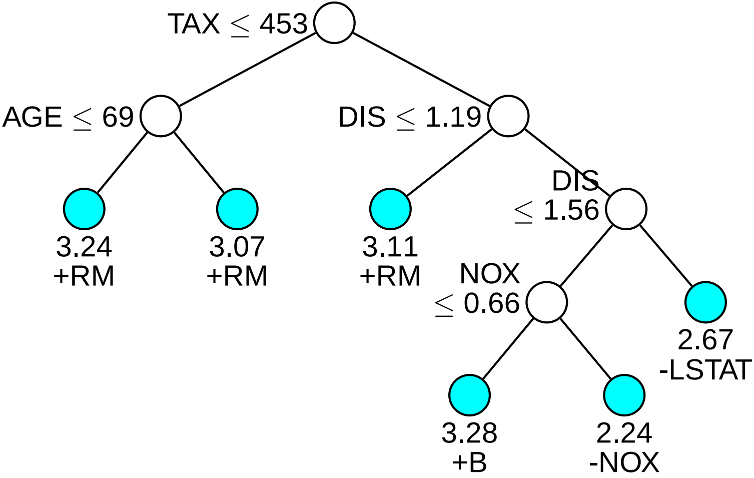
Why one- and two-predictor models?

- Data and model can be visualized with 2D and 3D plots
- Model can be interpreted unambiguously

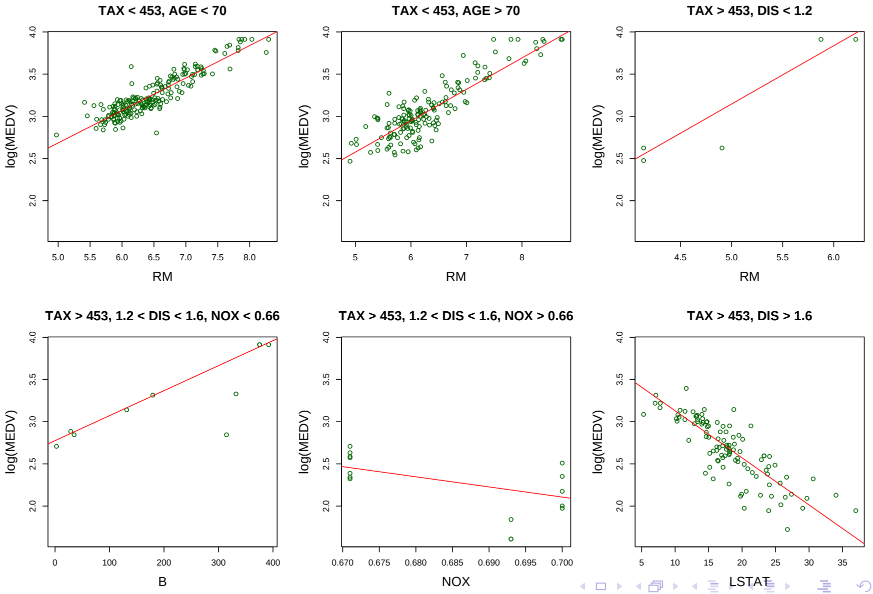
Least squares or robust?

- Robust fits are more resistant to outliers
- But robust fits may be less efficient

GUIDE robust regression model for log(MEDV)



Data and fits in terminal nodes



Resolving the conflict in the signs of log(DIS)

Resolving the conflict in the signs of log(DIS)

Strategy

1. Remove effects of other predictors by conditioning on them
2. Regress on log(DIS)

Resolving the conflict in the signs of log(DIS)

Strategy

1. Remove effects of other predictors by conditioning on them
2. Regress on log(DIS)

Procedure

- Make log(`DIS`) the only linear predictor
- Use all other predictors for splitting
- Let GUIDE construct the tree

Resolving the conflict in the signs of log(DIS)

Strategy

1. Remove effects of other predictors by conditioning on them
2. Regress on log(DIS)

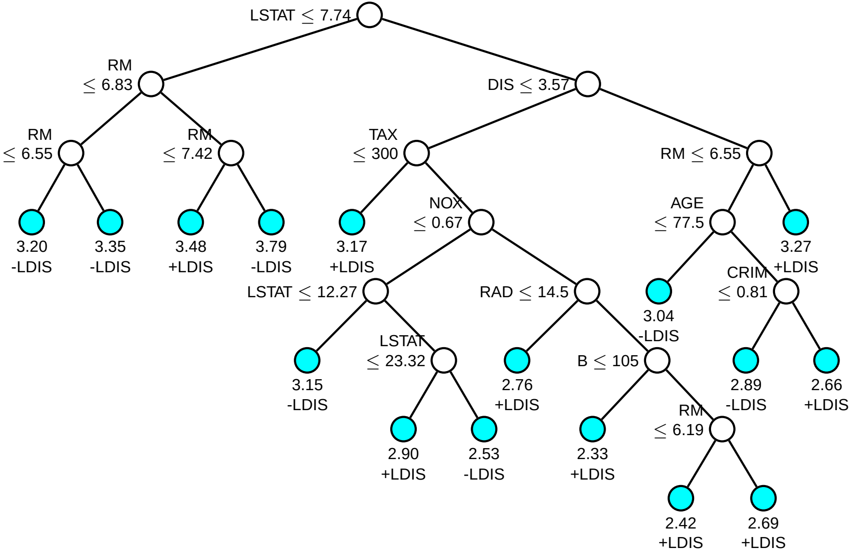
Procedure

- Make log(`DIS`) the only linear predictor
- Use all other predictors for splitting
- Let GUIDE construct the tree

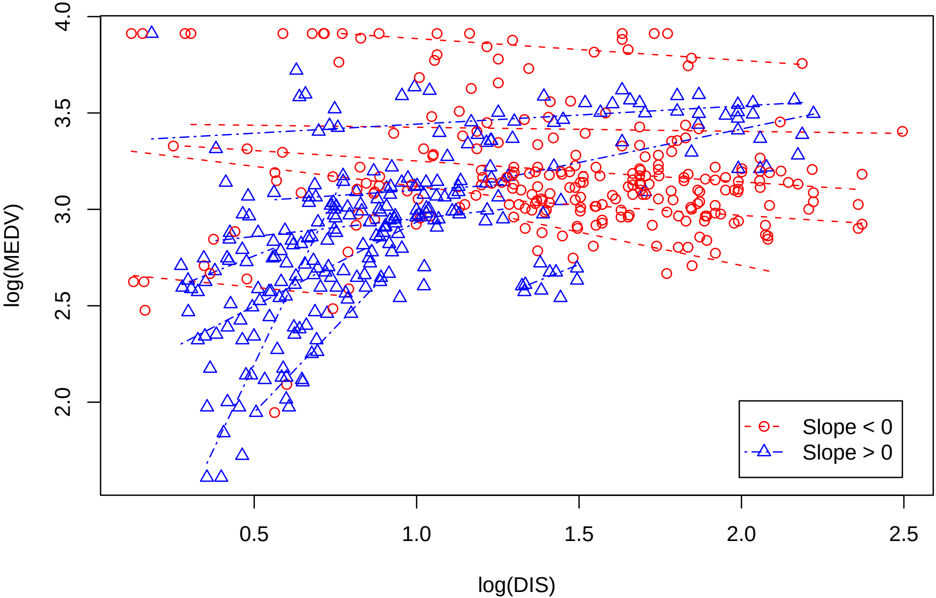
Advantage

No need to find a global model for log(MEDV)

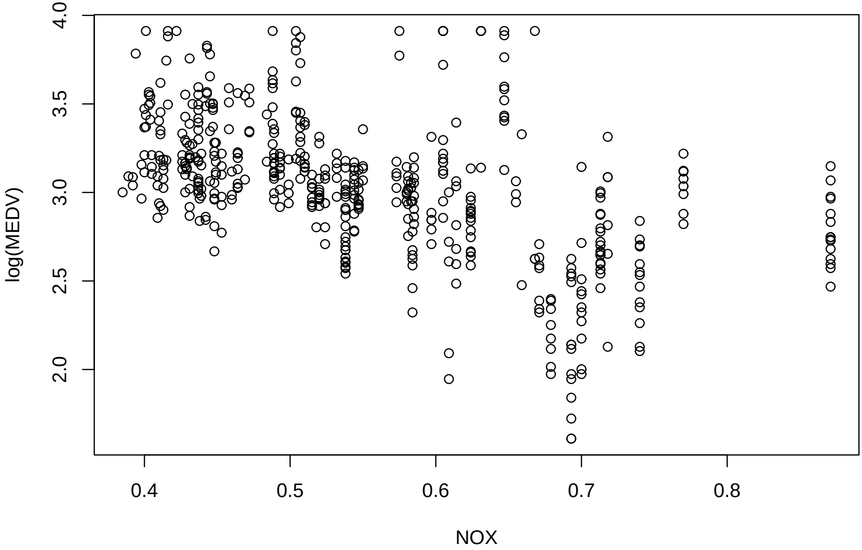
GUIDE model with log(DIS) as sole linear predictor



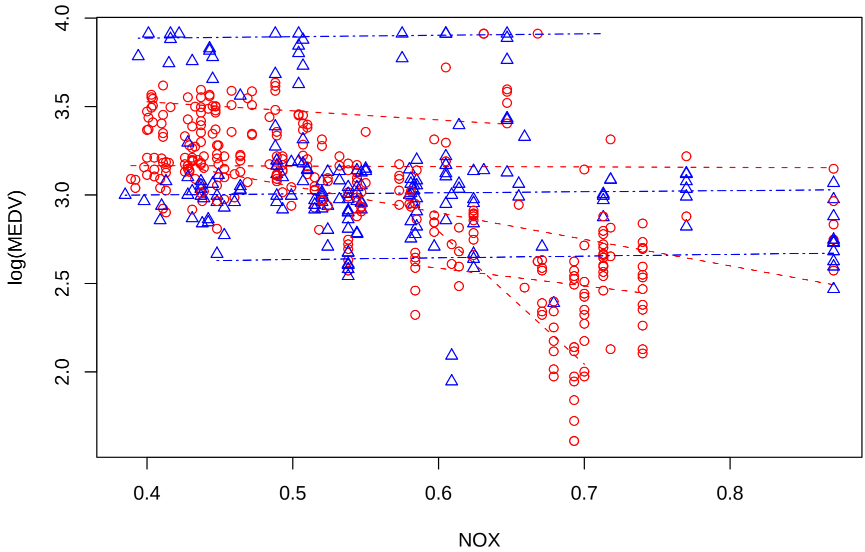
Data and fits in GUIDE model



The real question: what is the effect of NOX?



Effect of NOX after partitioning by GUIDE



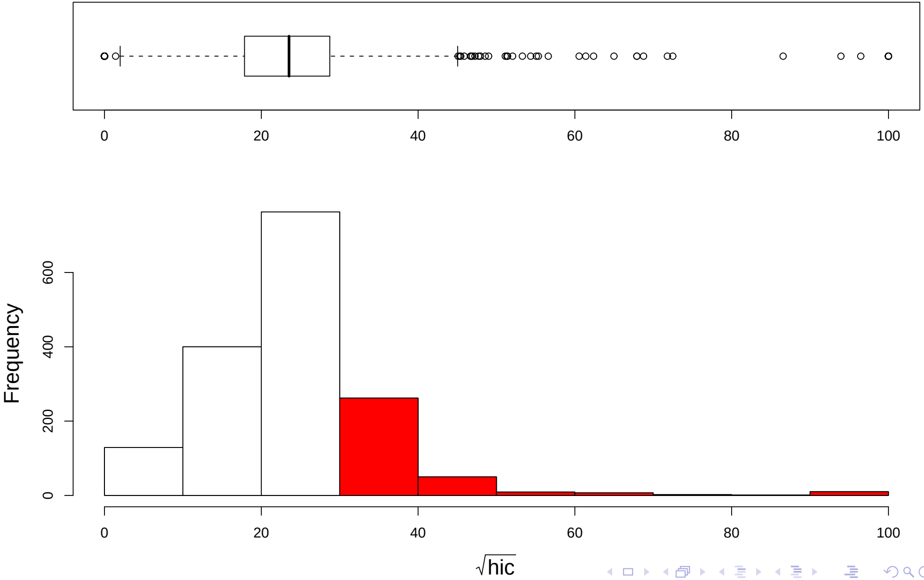
Application to outlier detection — vehicle crash tests

- National Highway Transportation Safety Administration (NHTSA) has been crash-testing vehicles since 1972
- 1,789 vehicles tested as of 2004
- One variable is head injury criterion (hic)
- $0 \le \sqrt{\texttt{hic}} < 100$
- Threshold for severe head injury is $\sqrt{\texttt{hic}} = 30$
- Twenty-five predictor variables give information on the vehicles, dummies, and test conditions

Our goal

Identify the vehicle models that are exceptionally unsafe (outliers) after controlling for the other variables

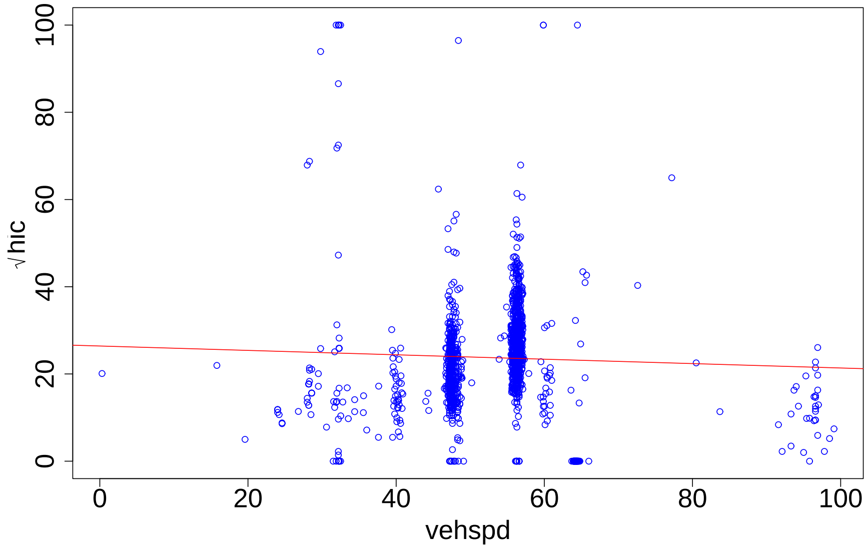
Boxplot and histogram for $\sqrt{\texttt{hic}}$ (driver data)



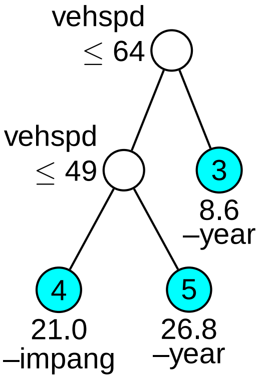
NHTSA variables (#distinct values in parentheses)

Name	Description	Name	Description
hic	Head injury criterion	make	Car manufacturer (62)
year	Car model year	mkmodel	Car model (464)
body	Car body type (18)	transm	Transmission type (7)
engine	Engine type (15)	engdsp	Engine displacement (liters)
vehtwt	Vehicle weight (kg)	colmec	Collapse mechanism (11)
vehwid	Vehicle width (mm)	modind	Modification indicator (5)
vehspd	Vehicle speed (km/h)	crbang	Crabbed angle
tksurf	Track surface (5)	pdof	Principal direction of force
tkcond	Track condition (6)	impang	Impact angle
occtyp	Occupant type (10)	dumsiz	Dummy size (6)
seposn	Seat position (5)	barrig	Barrier rigidity (2)
barshp	Barrier shape (14)	belts	Seat belt type (3)
airbag	Airbag present (2)	knee	Knee restraint present (2)

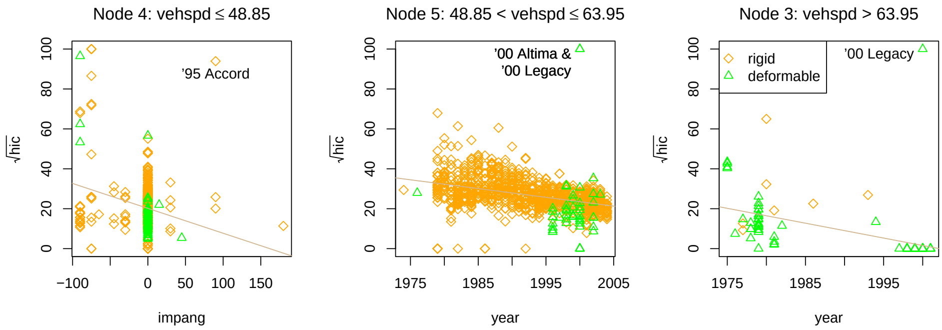
$\sqrt{\text{hic}}$ vs. vehspd — does speed kill?



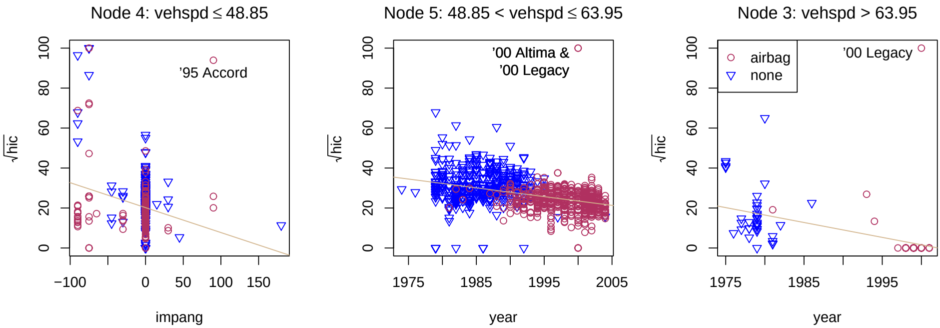
GUIDE piecewise linear model for $\sqrt{\texttt{hic}}$



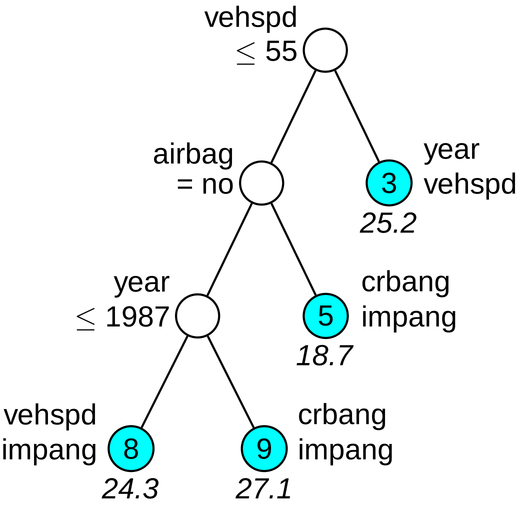
Data and fits in terminal nodes, by barrier rigidity



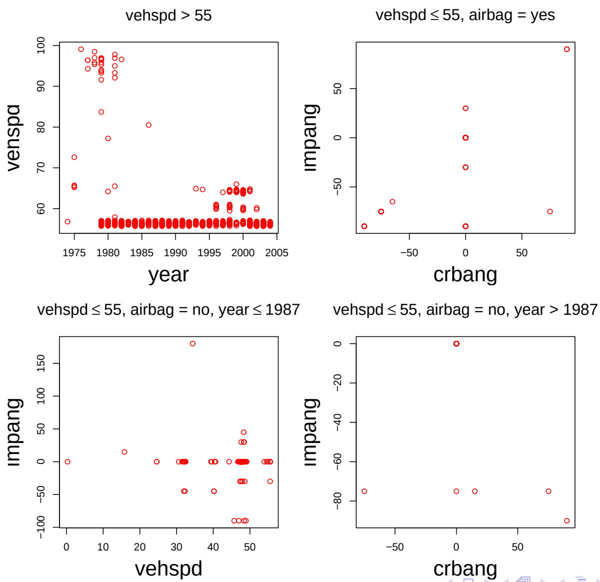
Data and fits in terminal nodes, by airbag



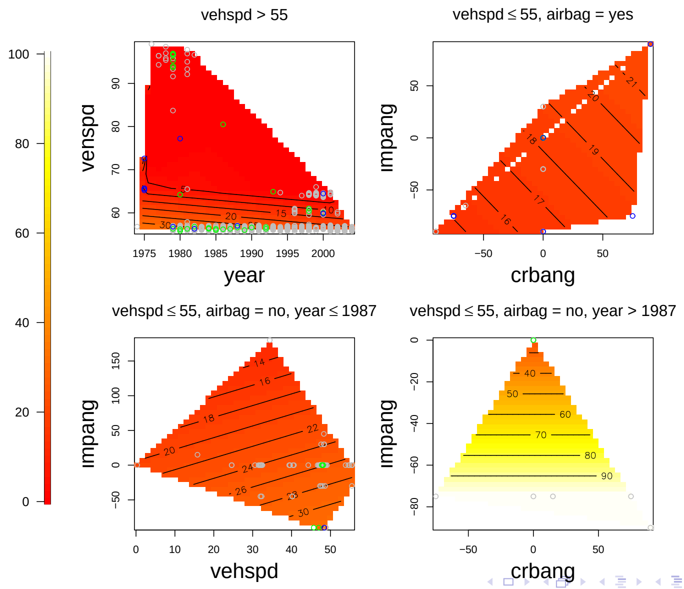
GUIDE piecewise two-predictor model for $\sqrt{\texttt{hic}}$



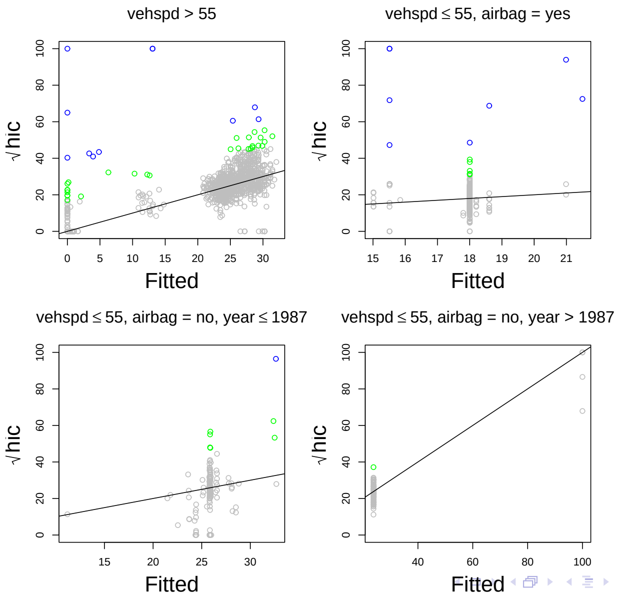
Distribution of data points



Contour plots of data and fitted functions



Blue points are 3 $\times$ *IQR* above 3rd quartile of residuals



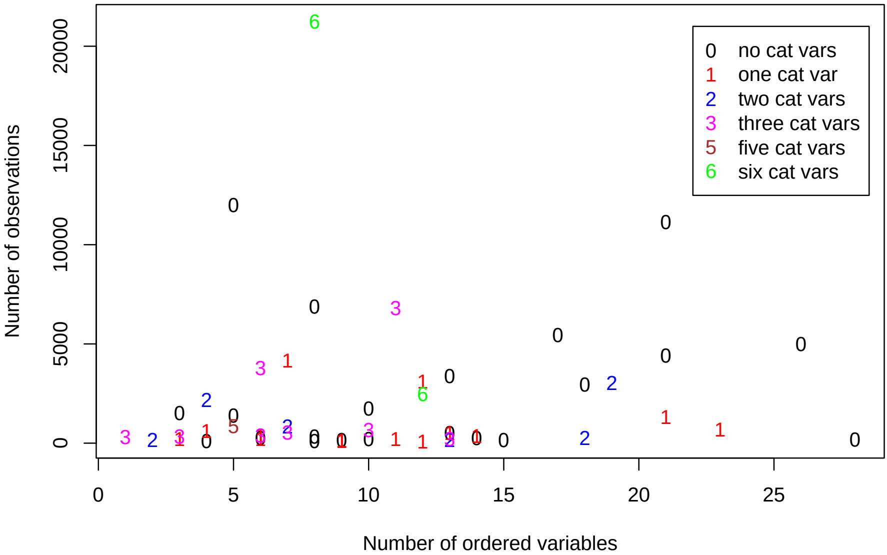
The most unsafe (blue) vehicles

1975 Ford Torino	1988 Chevy Sportvan
1975 Honda Civic	1988 Ford Tempo
1975 Plymouth Fury	1995 Honda Accord
1975 Volvo 244	2000 Nissan Altima
1979 Dodge Colt	2000 Nissan Maxima (3)
1979 Peugeot 504	2000 Saab 38235 (2)
1980 Chevy Citation	2000 Subaru Legacy (2)
1982 Renault Fuego	2002 Ford Explorer

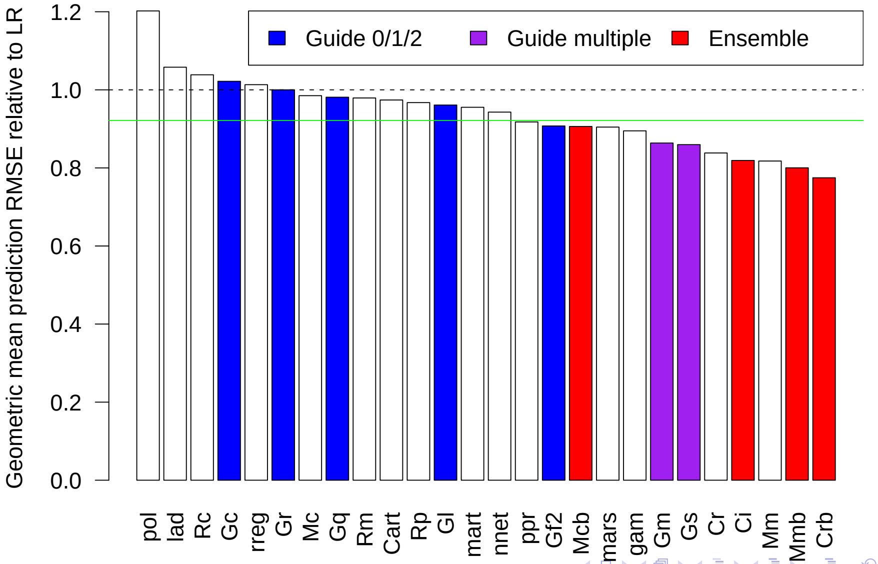
Prediction accuracy — 27 algorithms

Cart	CART	Mc	M5 constant
Cr	CUBIST rule-based	Mcb	Bagged Mc
Ci	CUBIST composite	Mm	M5 multiple linear
Crb	Boosted CUBIST	Mmb	Bagged Mm
Gc	GUIDE constant	mars	MARS
Gl	GUIDE simple linear	mart	MART
Gq	GUIDE simple quadratic	nnet	Neural network
Gm	GUIDE multiple linear	pol	POLYMARS
Gs	GUIDE stepwise linear	ppr	Projection pursuit
Gs2	GUIDE 2-regressor stepwise	Rc	RT constant
Gf2	GUIDE 2-regressor forward	Rm	RT multiple linear
gam	Generalized additive model	Rp	RT partial linear
lad	Least absolute deviation	rreg	Robust regression
lr	Least squares		

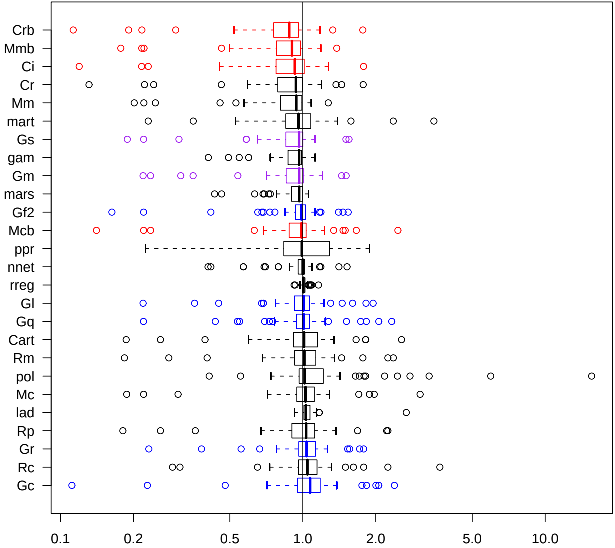
Prediction accuracy — 52 datasets



Prediction MSE relative to multiple linear regression



Boxplots of prediction MSE relative to LR



Concluding remarks

Advantages of piecewise simple and two-predictor models

1. Adaptive
2. Visualizable
3. Interpretable
4. At least as accurate as multiple linear regression

Concluding remarks

Advantages of piecewise simple and two-predictor models

1. Adaptive
2. Visualizable
3. Interpretable
4. At least as accurate as multiple linear regression

Future work

- Other robust methods (lower breakdown but higher efficiency)
- Robustify while-versus-after tree construction

Forthcoming papers

- Kim, Loh, Shih and Chaudhuri, “Visualizable and interpretable regression models with good prediction power.” Submitted to *IIE Transactions* Special Issue on Data Mining
- Loh, “Regression by parts: Fitting visually interpretable models with GUIDE.” To appear in *Handbook of Computational Statistics*, vol. III. Springer
- Loh, “Logistic regression tree analysis.” To appear in *Handbook of Engineering Statistics*. Springer

Download GUIDE:
`http://www.stat.wisc.edu/~loh/guide.html`