

Bias, Variance and Parsimony in Regression Analysis

ECS 256 Winter 2014

Christopher Patton, cjpatton@ucdavis.edu
Alex Rumbaugh, aprumbaugh@ucdavis.edu
Thomas Provan, tcprovan@ucdavis.edu
Olga Prilepova, prilepova@gmail.com
John Chen, jhochchen@ucdavis.edu

ECS 256, Winter 2014

UC Davis

March 12, 2014

Introduction

California Housing Data

- Derived from 1990 Census
- Response Variable: median house value
- Predictor Variables: median income, housing median age, total rooms, total bedrooms, population, households, latitude, and longitude

Parsimony

Method	Parsimony ($k=0.01$)	Parsimony ($k=0.05$)	Sig Test
Columns Deleted	Total Rooms Total Bedrooms	Total Rooms Total Bedrooms Median Age	None
Adjusted R^2	0.6321316	0.6218261	0.6369649

Regression Coefficients

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-3.594e+06	6.254e+04	-57.468	< 2e-16	***
Median.Income	4.025e+04	3.351e+02	120.123	< 2e-16	***
Median.Age	1.156e+03	4.317e+01	26.787	< 2e-16	***
Total.Rooms	-8.182e+00	7.881e-01	-10.381	< 2e-16	***
Total.Bedrooms	1.134e+02	6.902e+00	16.432	< 2e-16	***
Population	-3.854e+01	1.079e+00	-35.716	< 2e-16	***
Households	4.831e+01	7.515e+00	6.429	1.32e-10	***
Latitude	-4.258e+04	6.733e+02	-63.240	< 2e-16	***
Longitude	-4.282e+04	7.130e+02	-60.061	< 2e-16	***

Latitude & Longitude

Latitude	-4.258e+04	6.733e+02	-63.240	< 2e-16	***
Longitude	-4.282e+04	7.130e+02	-60.061	< 2e-16	***

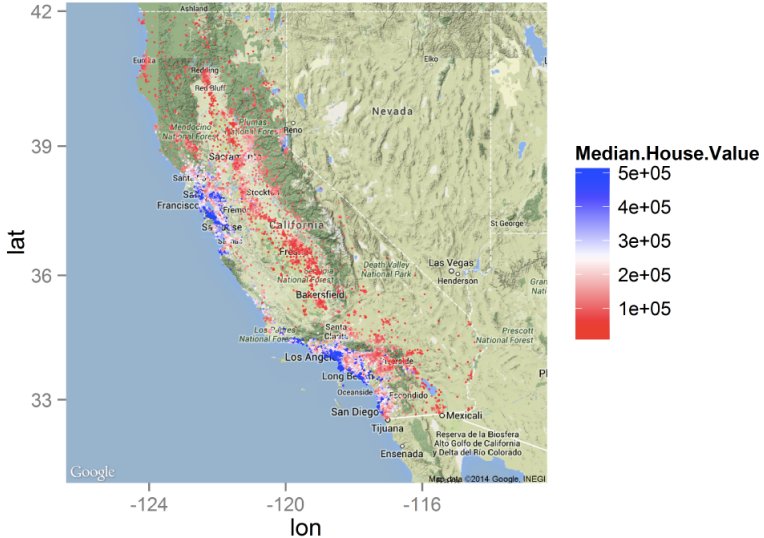
- "Center of Gravity"
- Avoid Overfitting

Understanding

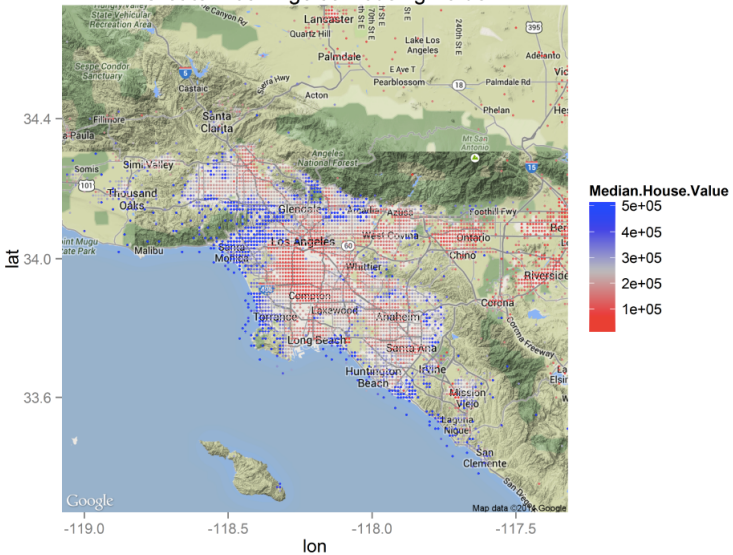
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-32165.268	2167.358	-14.84	<2e-16	***
Median.Income	43094.918	284.263	151.60	<2e-16	***
Median.Age	2000.544	45.080	44.38	<2e-16	***
Population	-43.045	1.127	-38.20	<2e-16	***
Households	152.700	3.344	45.66	<2e-16	***

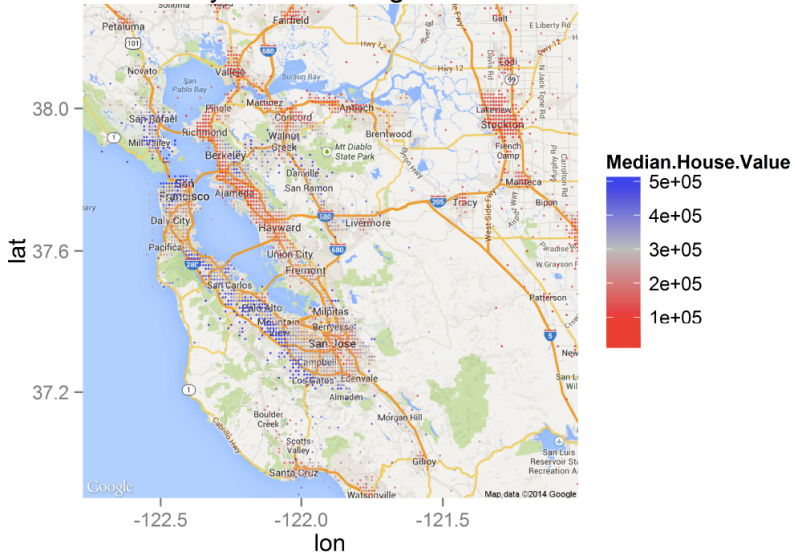
California Median House Value



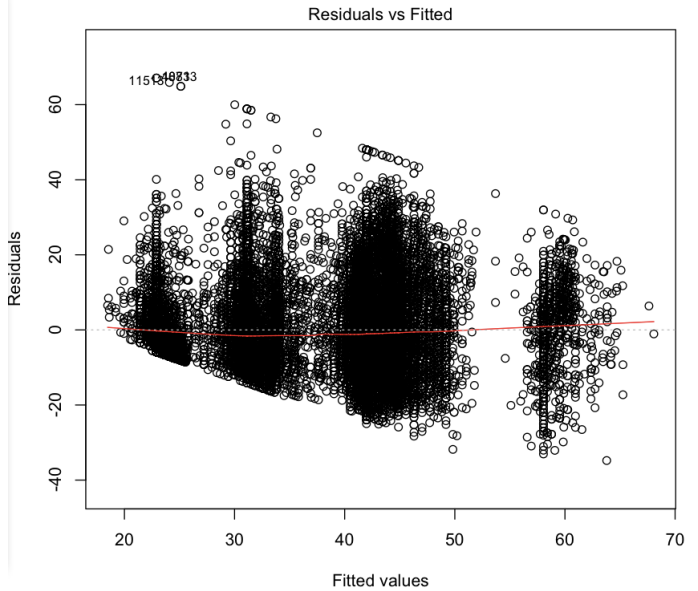
Greater Los Angeles Housing Value



Bay Area Housing Value



Census Based on 1994

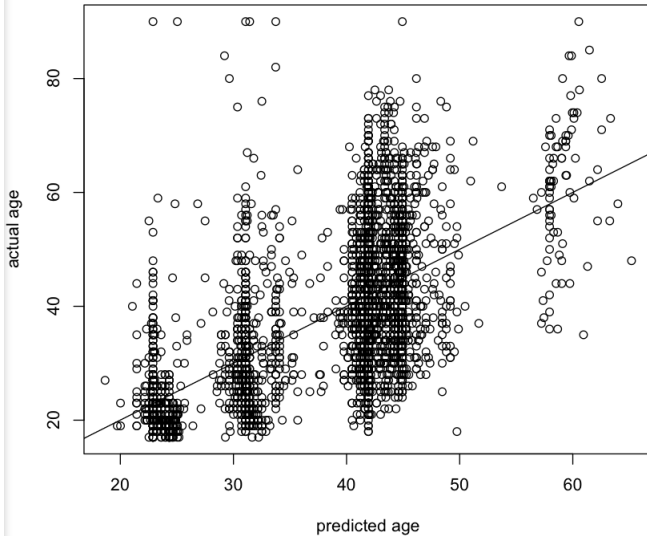


lm(age ~ Self.emp.not.inc + Assoc.acdm + Never.married + Widowed + Own.chil ...

Census Based on 1994

Census Based on 1994

predicted vs actual age of 15% test



PAC delta contributions

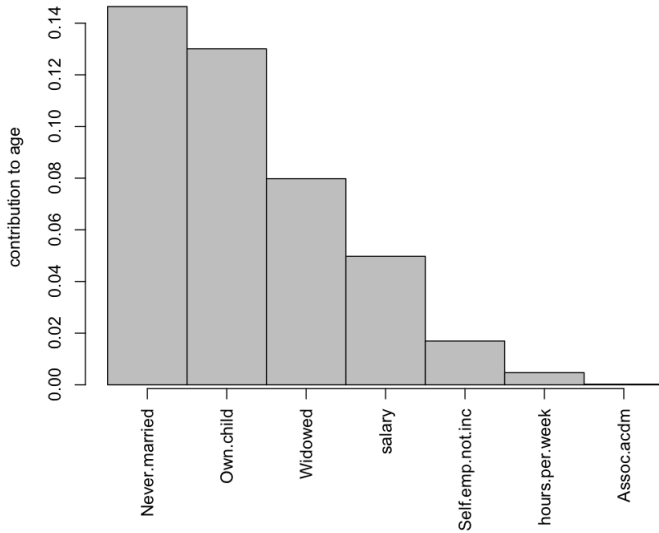


Figure:

Testing Parsimony on Simulated Data

Predictors: $X = X_1, \dots, X_{10}$

Response: Y drawn from $U(m_{Y;X}(t) - 1, m_{Y;X}(t) + 1)$

where $m_{Y;X}(t) = t_1 + t_2 + t_3 + 0.1t_4 + 0.01t_5$

Testing Parsimony on Simulated Data

		prsm(k=0.01)	prsm(k=0.05)	sig test
n=100	Run 1	X_1, X_2, X_3, X_9	X_1, X_2, X_3	X_1, X_2, X_3
	Run 2	X_1, X_2, X_3	X_1, X_2, X_3	X_1, X_2, X_3
	Run 3	X_1, X_2, X_3	X_1, X_2, X_3	X_1, X_2, X_3
n=1000	Run 1	X_1, X_2, X_3	X_1, X_2, X_3	X_1, X_2, X_3, X_4
	Run 2	X_1, X_2, X_3	X_1, X_2, X_3	X_1, X_2, X_3
	Run 3	X_1, X_2, X_3	X_1, X_2, X_3	X_1, X_2, X_3
n=10K	Run 1	X_1, X_2, X_3	X_1, X_2, X_3	X_1, X_2, X_3, X_4
	Run 2	X_1, X_2, X_3	X_1, X_2, X_3	X_1, X_2, X_3, X_4
	Run 3	X_1, X_2, X_3	X_1, X_2, X_3	X_1, X_2, X_3, X_4, X_9
n=100K	Run 1	X_1, X_2, X_3	X_1, X_2, X_3	X_1, X_2, X_3, X_4
	Run 2	X_1, X_2, X_3	X_1, X_2, X_3	X_1, X_2, X_3, X_4, X_9
	Run 3	X_1, X_2, X_3	X_1, X_2, X_3	X_1, X_2, X_3, X_4, X_9

Testing Parsimony on Simulated Data

k=0.01	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
N = 100	1	1	1	0.24	0.11	0.14	0.21	0.22	0.26	0.28
N = 1000	1	1	1	0.08	0	0	0	0	0	0
N = 10K	1	1	1	0	0	0	0	0	0	0
N = 100K	1	1	1	0	0	0	0	0	0	0
N = 1M	1	1	1	0	0	0	0	0	0	0
k=0.05	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
N = 100	1	1	0.99	0.1	0.02	0.05	0.04	0.03	0.07	0.02
N = 1000	1	1	1	0	0	0	0	0	0	0
N = 10K	1	1	1	0	0	0	0	0	0	0
N = 100K	1	1	1	0	0	0	0	0	0	0
N = 1M	1	1	1	0	0	0	0	0	0	0

Testing Parsimony on Simulated Data

Sig Test	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
N = 100	1	1	1	0.14	0.03	0.05	0.05	0.03	0.09	0.04
N = 1000	1	1	1	0.31	0.02	0.05	0.05	0.05	0.02	0.04
N = 10K	1	1	1	1	0.04	0.01	0.07	0.07	0.03	0.06
N = 100K	1	1	1	1	0.35	0.06	0.09	0.03	0.05	0.03
N = 1M	1	1	1	1	1	0.05	0.03	0.08	0.02	0.03

Small N, Large P

Automobile Data Set:

- UCI Machine Learning Repository
- 195 automobiles,
- 25 attributes per entry.

Goals:

- Determine accurate predictors of vehicle price.
- Gauge characteristics of safe automobiles.

Parsimony: Automobile Prices

- What factors best predict a vehicle's price?
- What are traits that increase price?
- What are the ones that decrease it?

Method	Parsimony ($k = 0.01$)	Parsimony ($k = 0.05$)	Significance Testing
Columns Retained	ohcv, twelve-cylinders, engine.size, stroke, compression.ratio, peak.rpm	engine.size	bmw, dodge, 'mercedes-benz', mitsubishi, plymouth, porsche, saab, std, front, wheel.base, length, width, height, curb.weight, dohc, ohc, engine.size, peak.rpm
AIC	0.8676842	0.7888274	0.9308

Significance Testing: Auto Prices

Results of Significance Testing (Auto Price):

(Intercept)	-4.234e+04	1.125e+04	-3.764	0.000229	***
bmw	9.290e+03	8.611e+02	10.788	< 2e-16	***
dodge	-1.504e+03	8.532e+02	-1.762	0.079785	.
'mercedes-benz'	6.644e+03	1.003e+03	6.625	4.17e-10	***
mitsubishi	-2.628e+03	7.331e+02	-3.585	0.000438	***
plymouth	-1.628e+03	8.881e+02	-1.833	0.068485	.
porsche	4.053e+03	2.238e+03	1.811	0.071936	.
saab	2.413e+03	1.028e+03	2.347	0.020043	*
std	-1.109e+03	5.129e+02	-2.162	0.031973	*
front	-1.275e+04	2.663e+03	-4.785	3.63e-06	***
wheel.base	1.141e+02	7.390e+01	1.544	0.124355	.
length	-7.918e+01	4.225e+01	-1.874	0.062586	.
width	7.652e+02	2.029e+02	3.772	0.000222	***
height	-1.377e+02	1.164e+02	-1.183	0.238332	.
curb.weight	3.781e+00	1.118e+00	3.381	0.000890	***
dohc	1.569e+03	8.067e+02	1.944	0.053451	.
ohc	8.531e+02	4.575e+02	1.865	0.063911	.
engine.size	7.733e+01	1.035e+01	7.470	3.74e-12	***
peak.rpm	1.522e+00	3.938e-01	3.864	0.000157	***

Multiple R-squared: 0.9373, Adjusted R-squared: 0.9308
F-statistic: 144.5 on 18 and 174 DF, p-value: < 2.2e-16

Top Predictors - Price

- Engine specifications, machinery
- Adds Value: Luxury Brands (BMW, Porsche)
- Reduces Value: Front-based Engine (Found in lower-end vehicles), economy brands (Mitsubishi, Plymouth)

Parsimony: Auto Safety

- Each auto is rated from -3 to 3 by insurers. -3 is safest, 3 is least safe.
- Use logistic regression to determine attributes of safe vehicles

Method	Parsimony ($k = 0.01$)	Parsimony ($k = 0.05$)	Significance Testing
Columns Retained	saab, toyota, volkswagen, turbo, two-doors, hatchback, sedan, 4wd, rwd, rear, wheel.base, length, width, height, curb.weight, l, ohc, ohcf, ohcv, five-cylinders, four-cylinders, three-cylinders, twelve-cylinders, engine.size, 2bbl, idi, mfi, mpfi, spdi, bore, stroke, compression.ratio, horsepower, peak.rpm, city.mpg, highway.mpg	saab, toyota, volkswagen, turbo, two-doors, hatchback, sedan, 4wd, rwd, rear, wheel.base, length, width, height, curb.weight, l, ohc, ohcf, ohcv, five-cylinders, four-cylinders, three-cylinders, twelve-cylinders, engine.size, 2bbl, idi, mfi, mpfi, spdi, bore, stroke, compression.ratio, horsepower, peak.rpm, city.mpg, highway.mpg	audi, saab, volkswagen, diesel, std, four-doors, 4wd, fwd, 1bbl
AIC	74	74	130.24

Significance Testing: Auto Safety

Results of Significance Testing (Auto Safety):

Coefficients:

	estimate	Std. Error	z value	Pr(> z)	
(Intercept)	E 2.5122	1.1216	2.240	0.02510	*
audi	20.3574	2027.3521	0.010	0.99199	
saab	17.7446	1985.9220	0.009	0.99287	
volkswagen	1.8112	0.9634	1.880	0.06011	.
diesel	-2.0155	1.2716	-1.585	0.11297	
std	-0.4196	1.0765	-0.390	0.69668	
'four-doors'	-5.9725	1.1293	-5.288	1.23e-07	***
'4wd'	-0.1377	2.1849	-0.063	0.94976	
fwd	3.3028	1.1093	2.977	0.00291	**
'1bbl'	-4.4965	1.4035	-3.204	0.00136	**

Null deviance: 266.06 on 192 degrees of freedom
Residual deviance: 110.24 on 183 degrees of freedom
AIC: 130.24

Top Predictors - Safety

- A negative z is a safer vehicle.
- The larger four-doored vehicles tend to be safer than two-doored ones.
- Sporty, rear-wheel drive vehicles tend to be more risky.
- `prsm()` unsuited for dimension reduction in this case - not enough data points. (Plymouth)

Bias, Variance and Parsimony in Regression Analysis

ECS 256 Winter 2014

Christopher Patton, cjpatton@ucdavis.edu

Alex Rumbaugh, aprumbaugh@ucdavis.edu

Thomas Provan, tcprovan@ucdavis.edu

Olga Prilepova, prilepova@gmail.com

John Chen, jhochchen@ucdavis.edu

ECS 256, Winter 2014

UC Davis

March 12, 2014