Boston Housing Data

Team 9

Seowoo Kim, Yun Jin Park, Taeim Kwon, Luyang You, Renee Ge

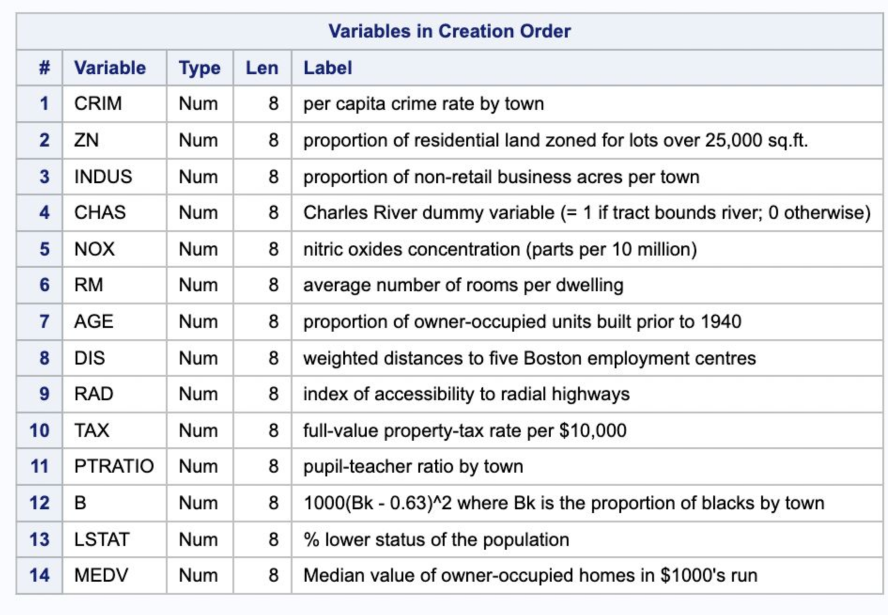
Motivation

What is the best model for predicting housing prices in Boston?

About the Dataset

- Each row represent one of Boston’s census tracts (suburb/town) from 1970
- 506 Observations
- Outcome of interest: MEDV
    - Median value of owner-occupied homes in the $1000s

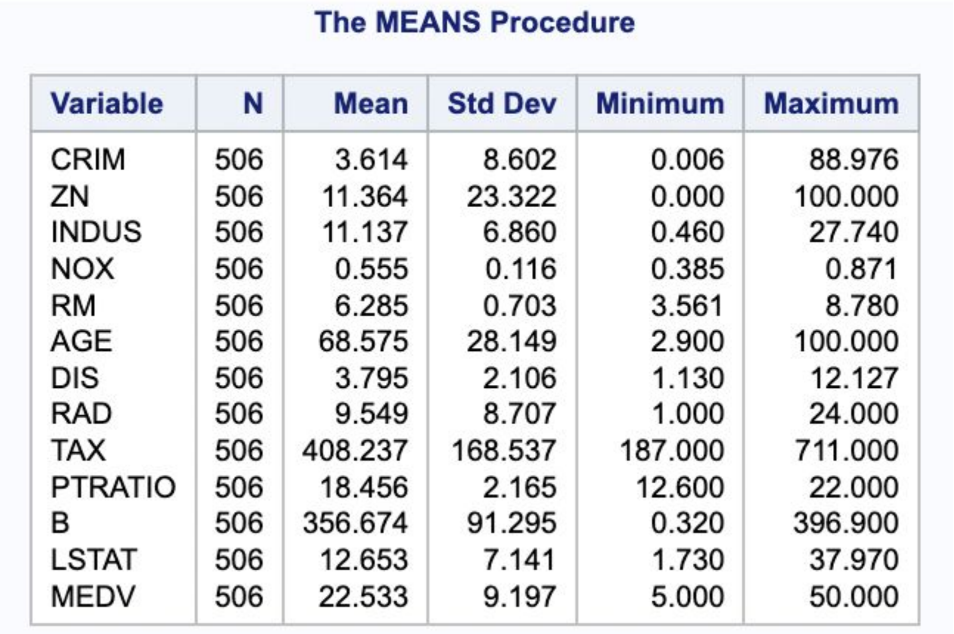
Summary Statistics



Variables in Creation Order

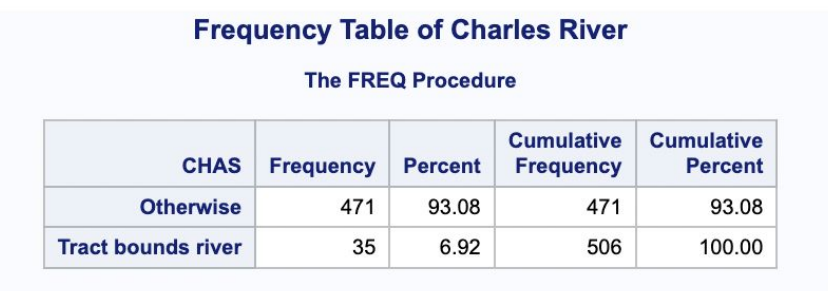
#	Variable	Type	Len	Label
1	CRIM	Num	8	per capita crime rate by town
2	ZN	Num	8	proportion of residential land zoned for lots over 25,000 sq.ft.
3	INDUS	Num	8	proportion of non-retail business acres per town
4	CHAS	Num	8	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5	NOX	Num	8	nitric oxides concentration (parts per 10 million)
6	RM	Num	8	average number of rooms per dwelling
7	AGE	Num	8	proportion of owner-occupied units built prior to 1940
8	DIS	Num	8	weighted distances to five Boston employment centres
9	RAD	Num	8	index of accessibility to radial highways
10	TAX	Num	8	full-value property-tax rate per $10,000
11	PTRATIO	Num	8	pupil-teacher ratio by town
12	B	Num	8	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13	LSTAT	Num	8	% lower status of the population
14	MEDV	Num	8	Median value of owner-occupied homes in $1000's run

Every variable except for CHAS is continuous.



The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
CRIM	506	3.614	8.602	0.006	88.976
ZN	506	11.364	23.322	0.000	100.000
INDUS	506	11.137	6.860	0.460	27.740
NOX	506	0.555	0.116	0.385	0.871
RM	506	6.285	0.703	3.561	8.780
AGE	506	68.575	28.149	2.900	100.000
DIS	506	3.795	2.106	1.130	12.127
RAD	506	9.549	8.707	1.000	24.000
TAX	506	408.237	168.537	187.000	711.000
PTRATIO	506	18.456	2.165	12.600	22.000
B	506	356.674	91.295	0.320	396.900
LSTAT	506	12.653	7.141	1.730	37.970
MEDV	506	22.533	9.197	5.000	50.000



Frequency Table of Charles River

The FREQ Procedure

CHAS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Otherwise	471	93.08	471	93.08
Tract bounds river	35	6.92	506	100.00

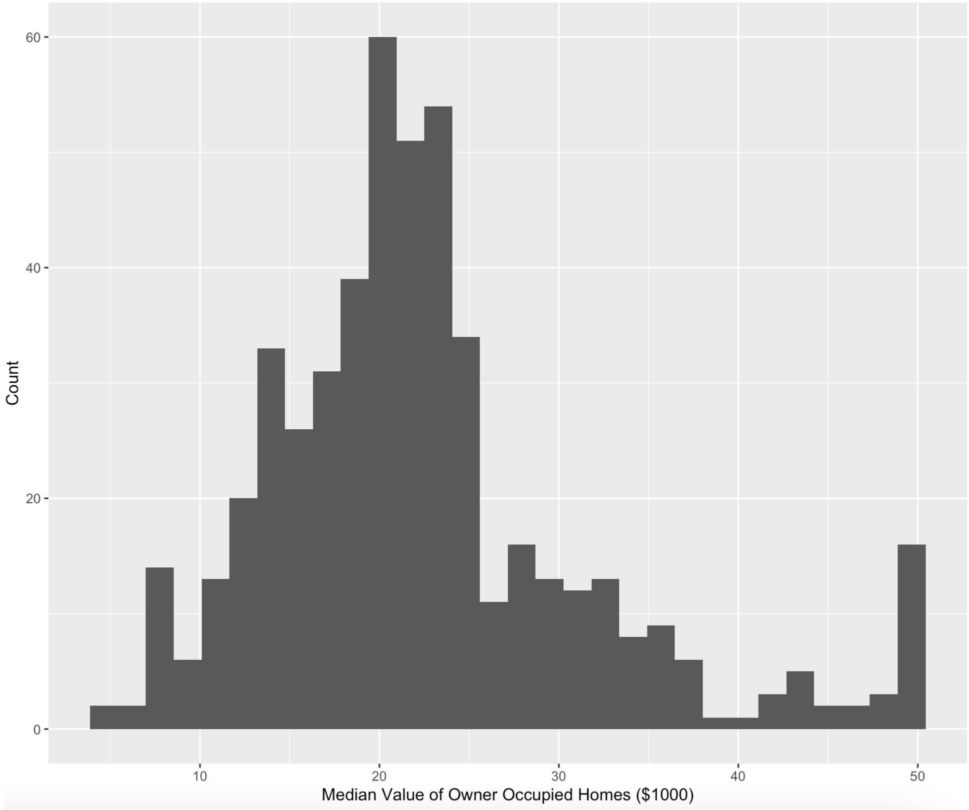
Analysis Plan

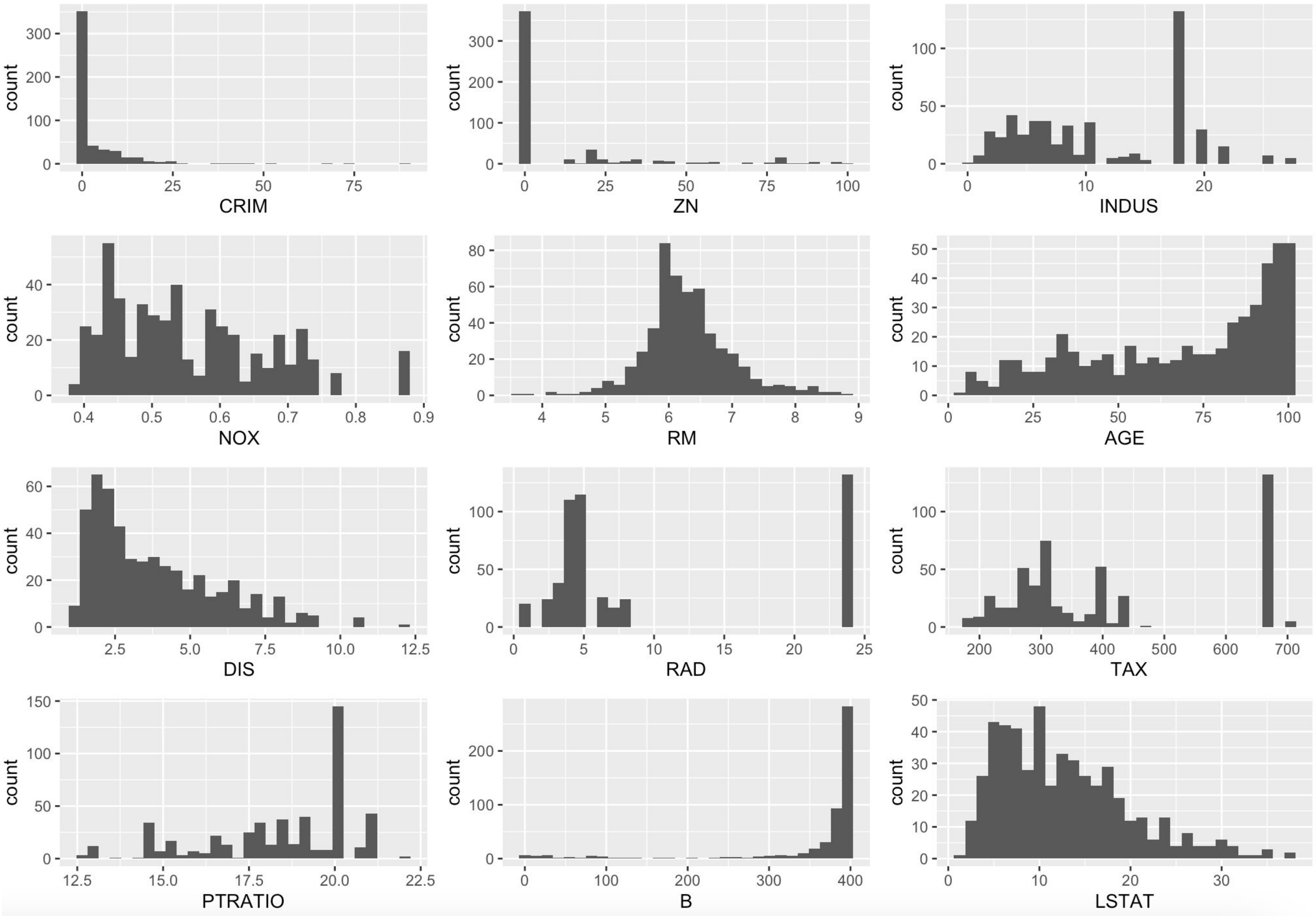
1. Exploratory Data Analysis

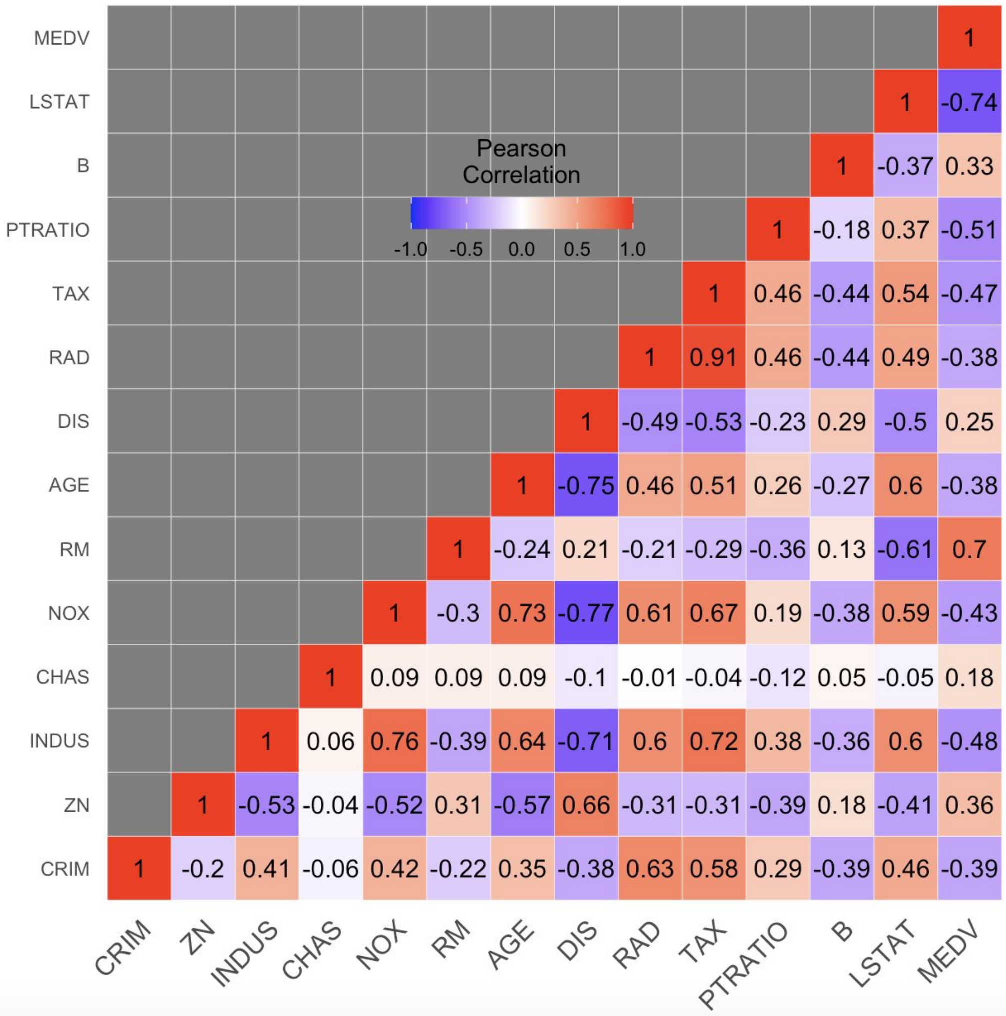
2. Fit Full model
   a. Perform Diagnostics
   b. Evaluate influential points
   c. Examine collinearity in data

3. Backwards Stepwise Regression with $\alpha = 0.05$

MEDV Outcome:




count
300
200
100
0
0
25
50
75
CRIM
count
300
200
100
0
0
25
50
75
100
ZN
count
100
50
0
0
10
20
INDUS
count
40
20
0
0.4
0.5
0.6
0.7
0.8
0.9
NOX
count
80
60
40
20
0
4
5
6
7
8
9
RM
count
50
40
30
20
10
0
0
25
50
75
100
AGE
count
60
40
20
0
2.5
5.0
7.5
10.0
12.5
DIS
count
100
50
0
0
5
10
15
20
25
RAD
count
100
50
0
200
300
400
500
600
700
TAX
count
150
100
50
0
12.5
15.0
17.5
20.0
22.5
PTRATIO
count
200
100
0
0
100
200
300
400
B
count
50
40
30
20
10
0
0
10
20
30
LSTAT

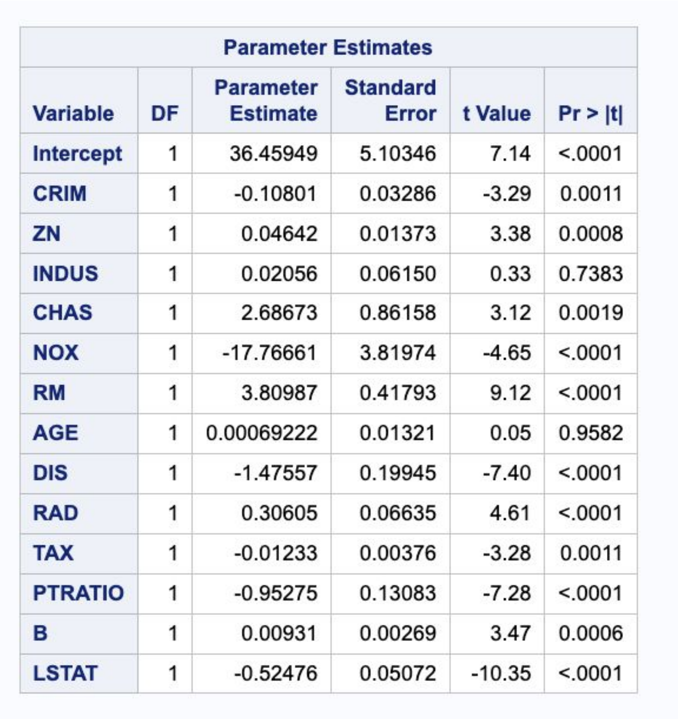


	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
MEDV														1
LSTAT													1	-0.74
B												1	-0.37	0.33
PTRATIO											1	-0.18	0.37	-0.51
TAX										1	0.46	-0.44	0.54	-0.47
RAD									1	0.91	0.46	-0.44	0.49	-0.38
DIS								1	-0.49	-0.53	-0.23	0.29	-0.5	0.25
AGE							1	-0.75	0.46	0.51	0.26	-0.27	0.6	-0.38
RM						1	-0.24	0.21	-0.21	-0.29	-0.36	0.13	-0.61	0.7
NOX					1	-0.3	0.73	-0.77	0.61	0.67	0.19	-0.38	0.59	-0.43
CHAS				1	0.09	0.09	0.09	-0.1	-0.01	-0.04	-0.12	0.05	-0.05	0.18
INDUS			1	0.06	0.76	-0.39	0.64	-0.71	0.6	0.72	0.38	-0.36	0.6	-0.48
ZN		1	-0.53	-0.04	-0.52	0.31	-0.57	0.66	-0.31	-0.31	-0.39	0.18	-0.41	0.36
CRIM	1	-0.2	0.41	-0.06	0.42	-0.22	0.35	-0.38	0.63	0.58	0.29	-0.39	0.46	-0.39

Full Model:

$$MEDV \ = \beta_0 + \sum_{i=1}^{13} \beta_i \, x_i + \varepsilon_i$$

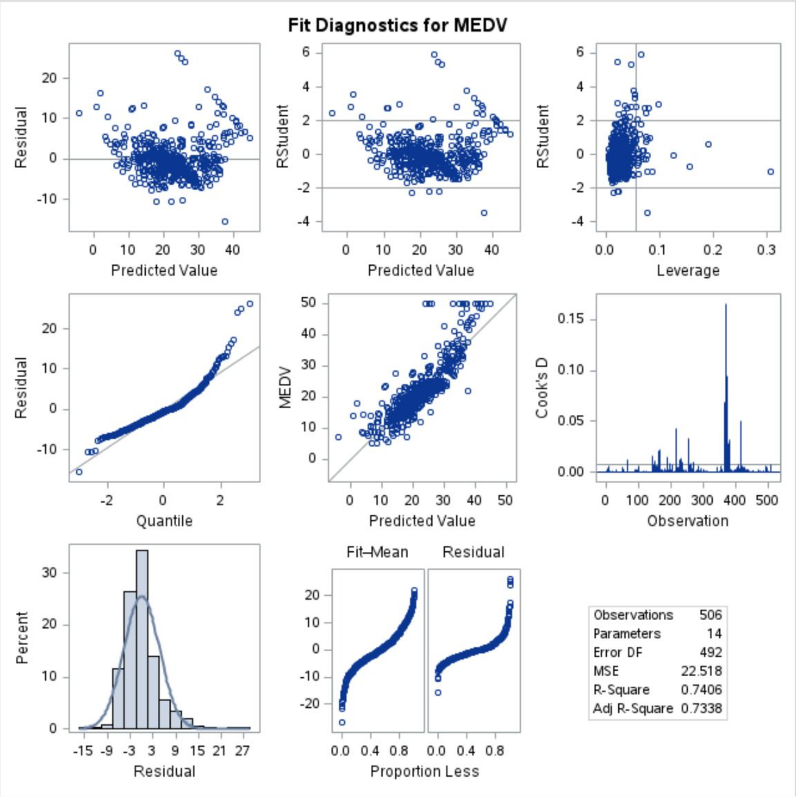
Full Model



Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	36.45949	5.10346	7.14	<.0001
CRIM	1	-0.10801	0.03286	-3.29	0.0011
ZN	1	0.04642	0.01373	3.38	0.0008
INDUS	1	0.02056	0.06150	0.33	0.7383
CHAS	1	2.68673	0.86158	3.12	0.0019
NOX	1	-17.76661	3.81974	-4.65	<.0001
RM	1	3.80987	0.41793	9.12	<.0001
AGE	1	0.00069222	0.01321	0.05	0.9582
DIS	1	-1.47557	0.19945	-7.40	<.0001
RAD	1	0.30605	0.06635	4.61	<.0001
TAX	1	-0.01233	0.00376	-3.28	0.0011
PTRATIO	1	-0.95275	0.13083	-7.28	<.0001
B	1	0.00931	0.00269	3.47	0.0006
LSTAT	1	-0.52476	0.05072	-10.35	<.0001

- AGE and INDUS have p-value > 0.05
- Adjusted R2 = 0.7338

Full model: Perform Diagnostics

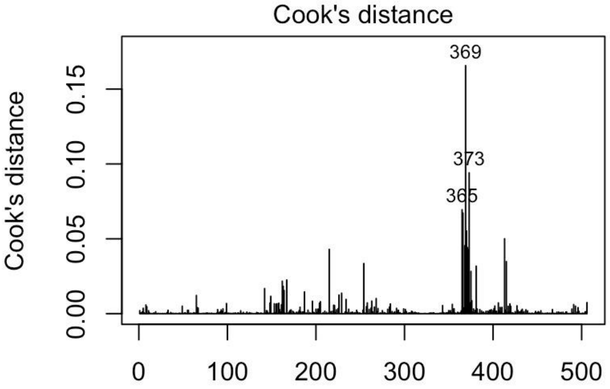


Observations	506
Parameters	14
Error DF	492
MSE	22.518
R-Square	0.7406
Adj R-Square	0.7338

- Existence assumption
- Independence assumption
- Linearity assumption
- Homogeneity assumption
- Gaussian errors assumption

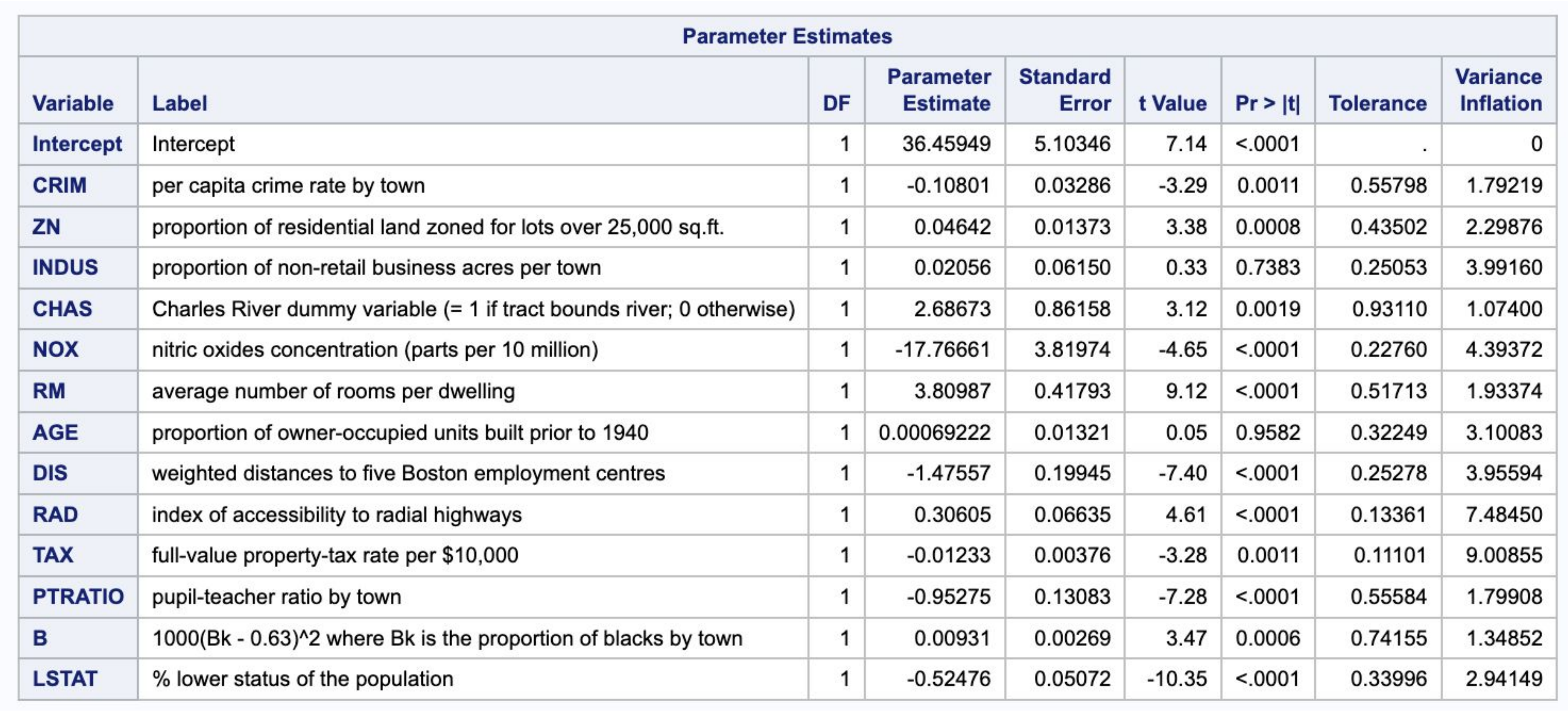
```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  model$residuals
D = 0.34781, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Full model: Evaluate influential points



Variable Names	365th Obs	369th Obs	373th Obs	Comment
CRIM	-0.016	0.149	0.541	
ZN	-0.487	-0.487	-0.487	
INDUS	1.015	1.015	1.015	
CHAS	3.665	-0.272	3.665	Categorical
NOX	1.409	0.658	0.978	
RM	3.552	-1.871	-0.583	365th Obs > 3 SD
AGE	0.509	1.116	0.747	
DIS	-0.898	-1.169	-1.266	
RAD	1.660	1.660	1.660	
TAX	1.529	1.529	1.529	
PTRATIO	0.806	0.806	0.806	
B	-0.023	0.206	-0.096	
LSTAT	-1.031	-1.315	-0.528	
MEDV	0.127	2.987	2.987	369th Obs >2 SD 373th Obs >2 SD

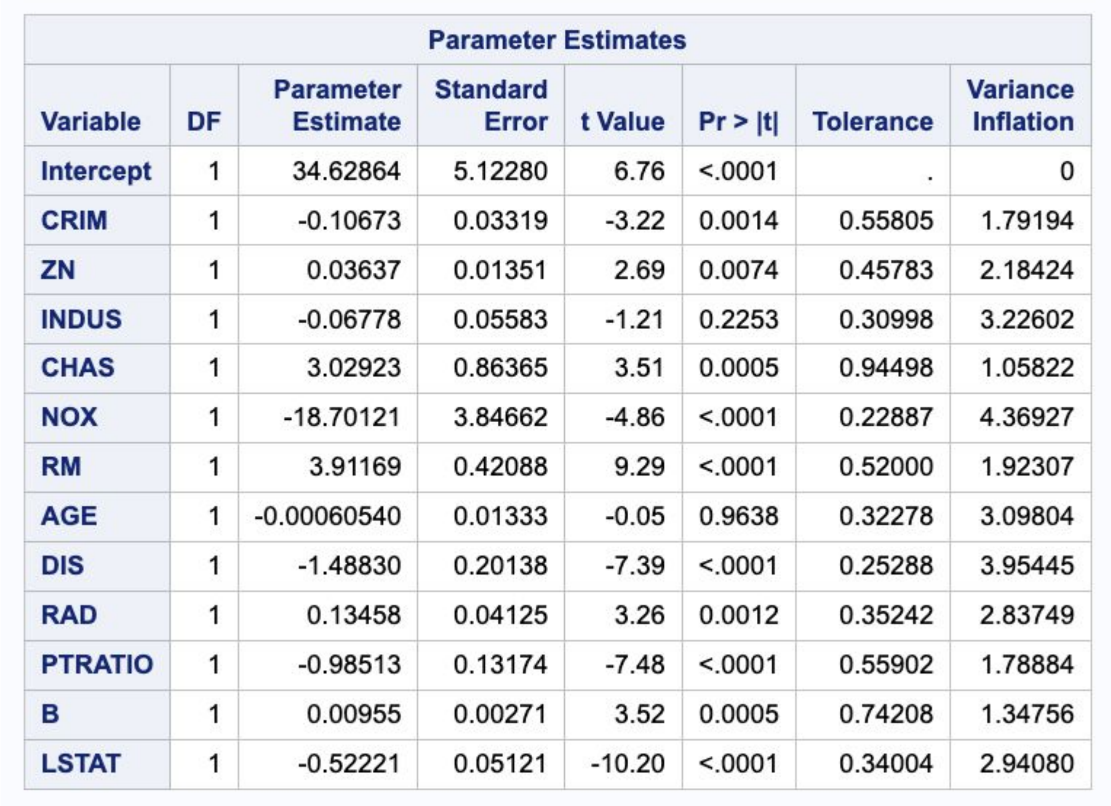
Check Collinearity (VIF)



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	36.45949	5.10346	7.14	<.0001	.	0
CRIM	per capita crime rate by town	1	-0.10801	0.03286	-3.29	0.0011	0.55798	1.79219
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.	1	0.04642	0.01373	3.38	0.0008	0.43502	2.29876
INDUS	proportion of non-retail business acres per town	1	0.02056	0.06150	0.33	0.7383	0.25053	3.99160
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)	1	2.68673	0.86158	3.12	0.0019	0.93110	1.07400
NOX	nitric oxides concentration (parts per 10 million)	1	-17.76661	3.81974	-4.65	<.0001	0.22760	4.39372
RM	average number of rooms per dwelling	1	3.80987	0.41793	9.12	<.0001	0.51713	1.93374
AGE	proportion of owner-occupied units built prior to 1940	1	0.00069222	0.01321	0.05	0.9582	0.32249	3.10083
DIS	weighted distances to five Boston employment centres	1	-1.47557	0.19945	-7.40	<.0001	0.25278	3.95594
RAD	index of accessibility to radial highways	1	0.30605	0.06635	4.61	<.0001	0.13361	7.48450
TAX	full-value property-tax rate per $10,000	1	-0.01233	0.00376	-3.28	0.0011	0.11101	9.00855
PTRATIO	pupil-teacher ratio by town	1	-0.95275	0.13083	-7.28	<.0001	0.55584	1.79908
B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town	1	0.00931	0.00269	3.47	0.0006	0.74155	1.34852
LSTAT	% lower status of the population	1	-0.52476	0.05072	-10.35	<.0001	0.33996	2.94149

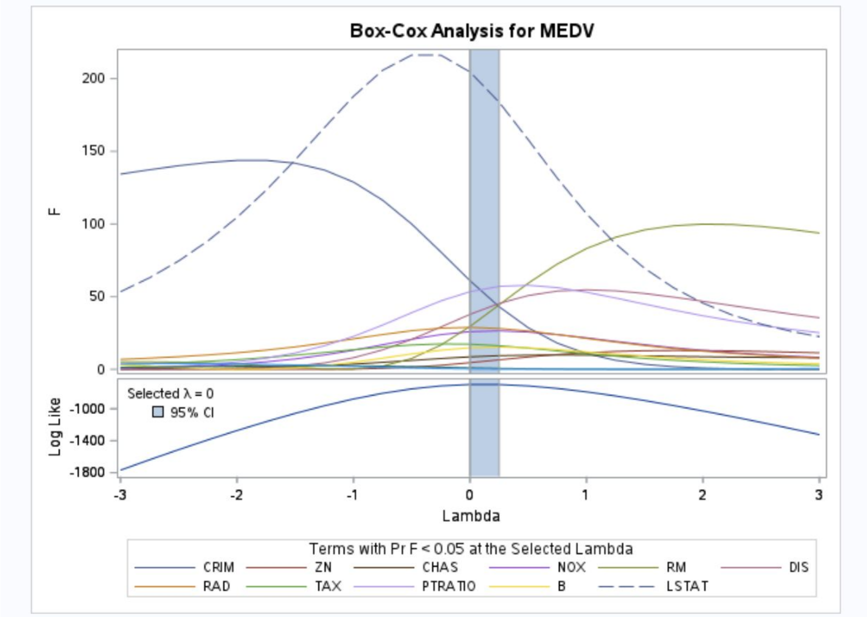
Check Collinearity (VIF)

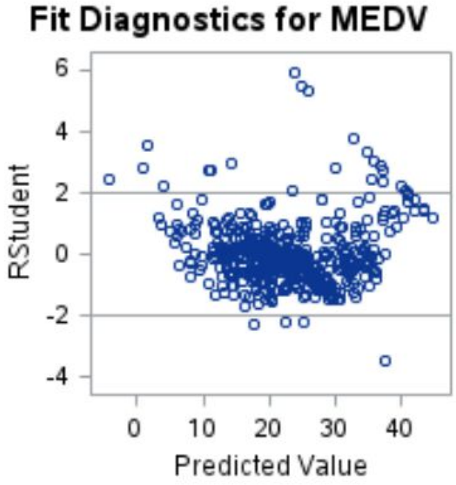


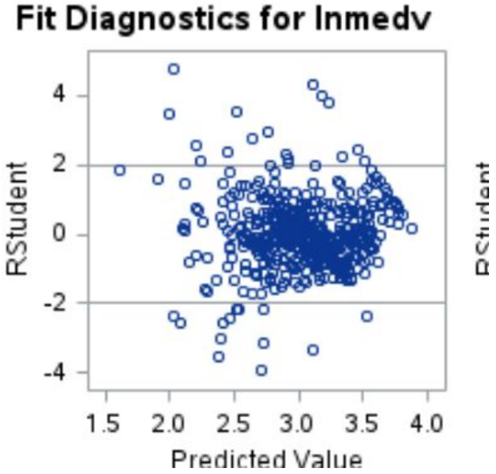
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	34.62864	5.12280	6.76	<.0001	.	0
CRIM	1	-0.10673	0.03319	-3.22	0.0014	0.55805	1.79194
ZN	1	0.03637	0.01351	2.69	0.0074	0.45783	2.18424
INDUS	1	-0.06778	0.05583	-1.21	0.2253	0.30998	3.22602
CHAS	1	3.02923	0.86365	3.51	0.0005	0.94498	1.05822
NOX	1	-18.70121	3.84662	-4.86	<.0001	0.22887	4.36927
RM	1	3.91169	0.42088	9.29	<.0001	0.52000	1.92307
AGE	1	-0.00060540	0.01333	-0.05	0.9638	0.32278	3.09804
DIS	1	-1.48830	0.20138	-7.39	<.0001	0.25288	3.95445
RAD	1	0.13458	0.04125	3.26	0.0012	0.35242	2.83749
PTRATIO	1	-0.98513	0.13174	-7.48	<.0001	0.55902	1.78884
B	1	0.00955	0.00271	3.52	0.0005	0.74208	1.34756
LSTAT	1	-0.52221	0.05121	-10.20	<.0001	0.34004	2.94080

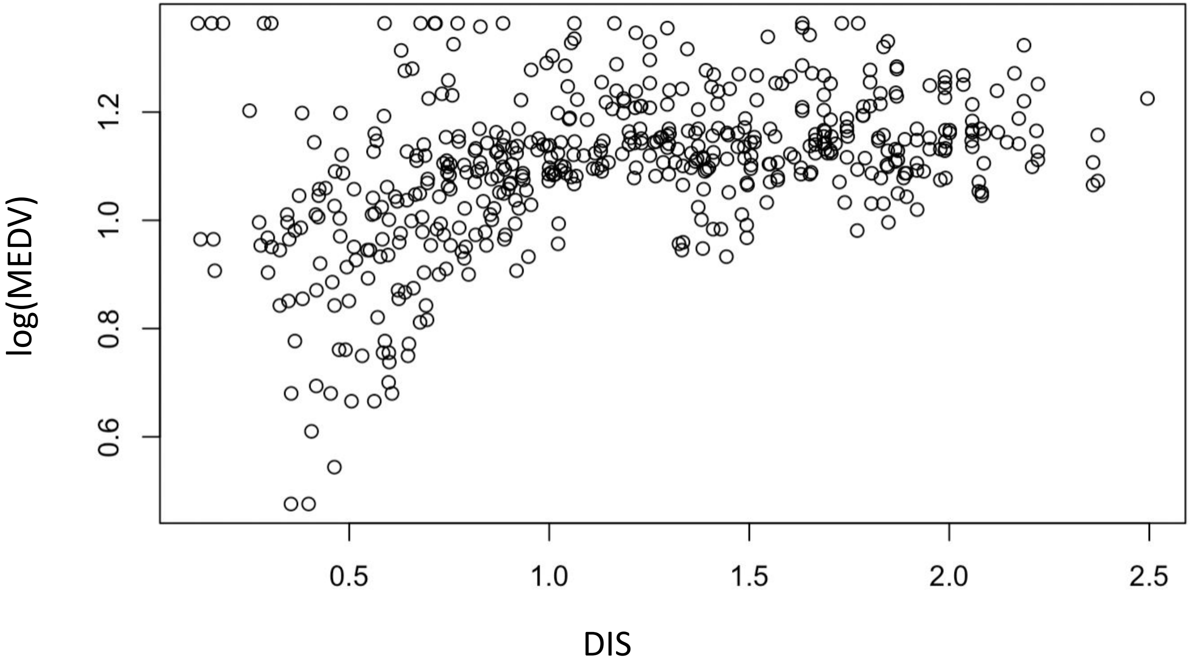
Box-Cox Transformation







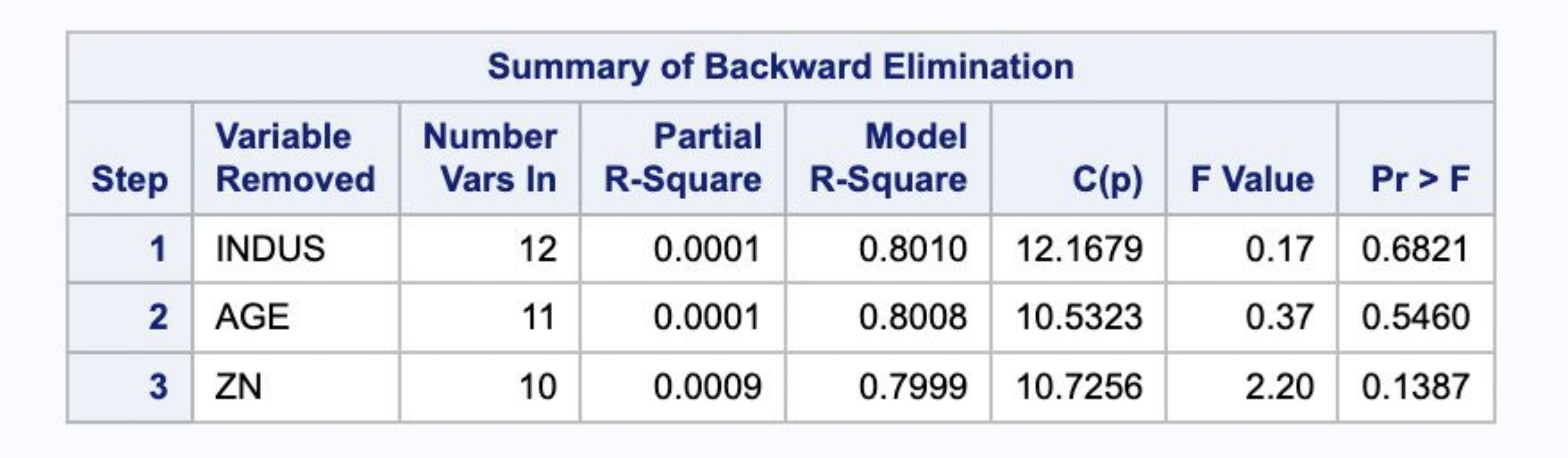
Predictor Transformation



The scatterplot shows a trend of log function

Backward Selection

Predictors “ZN(proportion of residential land zoned for lots over 25,000 sq.ft.)”, “AGE (proportion of owner-occupied units built prior to 1940)”, and “INDUS (proportion of non-retail business acres per town)” are eliminated.(significance level = 0.05)

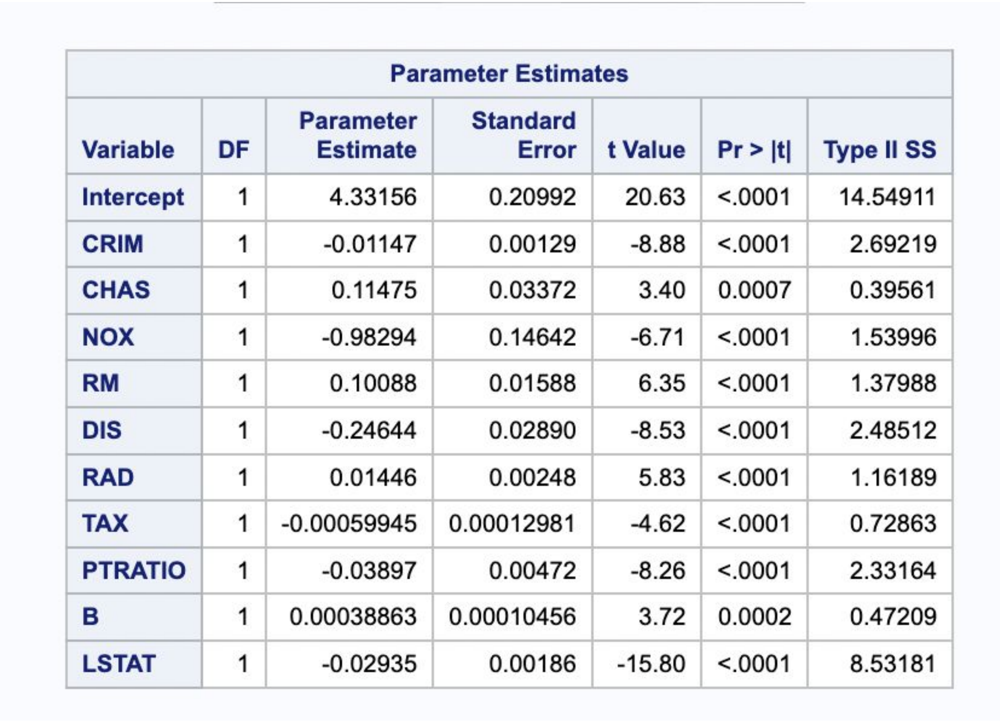


Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	INDUS	12	0.0001	0.8010	12.1679	0.17	0.6821
2	AGE	11	0.0001	0.8008	10.5323	0.37	0.5460
3	ZN	10	0.0009	0.7999	10.7256	2.20	0.1387

New Model:

$\log(\mathrm{MEDV}) = \beta_0 + \beta_1 \mathrm{CRIM} + \beta_2\ \mathrm{CHAS} + \beta_3\ \mathrm{NOX} + \beta_4\ \mathrm{RM} + \beta_5 \log(\mathrm{DIS}) + \beta_6\ \mathrm{RAD} + \beta_7\ \mathrm{PTRATIO} + \beta_8 \mathrm{B} + \beta_9\ \mathrm{LSTAT}$



Parameter Estimates

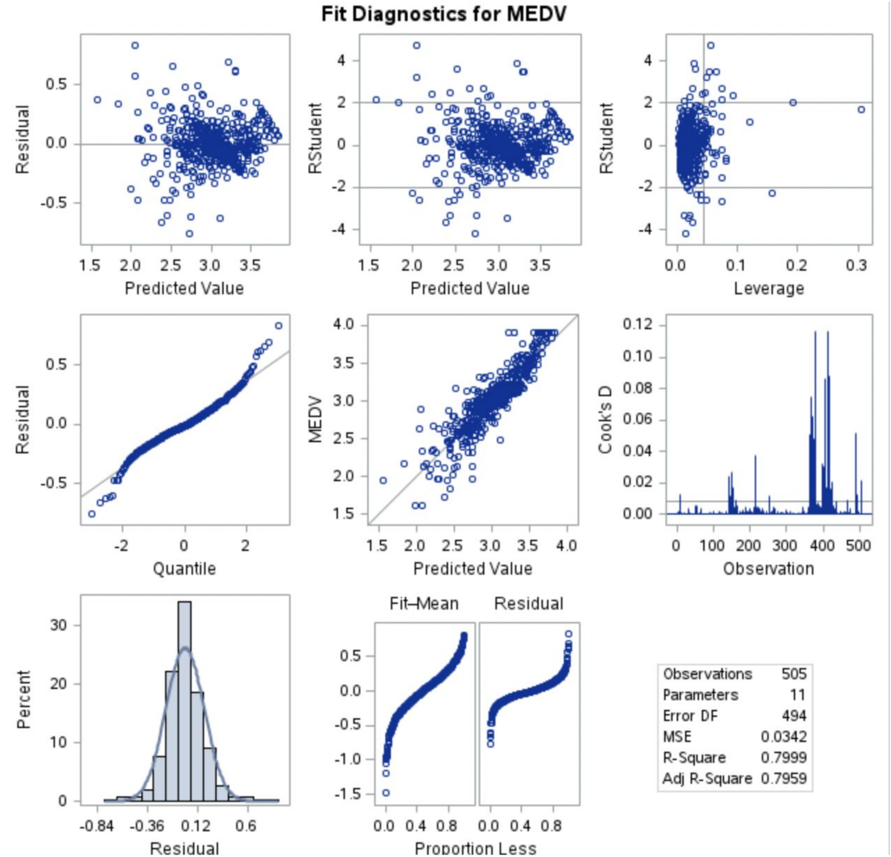
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type II SS
Intercept	1	4.33156	0.20992	20.63	<.0001	14.54911
CRIM	1	-0.01147	0.00129	-8.88	<.0001	2.69219
CHAS	1	0.11475	0.03372	3.40	0.0007	0.39561
NOX	1	-0.98294	0.14642	-6.71	<.0001	1.53996
RM	1	0.10088	0.01588	6.35	<.0001	1.37988
DIS	1	-0.24644	0.02890	-8.53	<.0001	2.48512
RAD	1	0.01446	0.00248	5.83	<.0001	1.16189
TAX	1	-0.00059945	0.00012981	-4.62	<.0001	0.72863
PTRATIO	1	-0.03897	0.00472	-8.26	<.0001	2.33164
B	1	0.00038863	0.00010456	3.72	0.0002	0.47209
LSTAT	1	-0.02935	0.00186	-15.80	<.0001	8.53181

Adjusted R2: 0.7959

Adjusted R2 from full model: 0.7338

Assumption Checking

- Existence assumption
- Independence assumption
- Linearity assumption
- Homogeneity assumption
- Gaussian errors assumption



Observations	505
Parameters	11
Error DF	494
MSE	0.0342
R-Square	0.7999
Adj R-Square	0.7959

Discussion

Covariates that have positive effects on log of median value of homes in Boston on average:

- Closer to Charles River
- Larger number of rooms per dwelling
- More accessible to radial highway
- More population homogeneity

Covariates that have negative effects on log of median value of homes in Boston on average:

- More crimes
- More air pollution
- Far from five Boston employment centres
- Higher tax rate
- Higher pupil-teacher ratio (lower quality of education)
- Higher proportion of lower status

Next steps

- Polynomial Regression

- Incorporating more recent data would give more accurate predictions
  - This data is from the 1970s, likely would not be applicable nowadays even in Boston

Thank you!