

# Project 1

Manager: Brendan McDonnel  
Masum Billah  
Madhu Chencharapu  
Gabriel Loos  
Roshan Ravichandran

February 2026

## 1 Introduction

## 2 Linear Regression

### 2.1 Auto MPG

Table 1: Scalation - AutoMPG Linear Regression

Metric	In-Sample	80-20 Split
rSq	0.809255	0.822842
rSqBar	0.806283	0.820081
sst	23819.0	4731.23
sse	4543.35	838.174
sde	3.40878	3.29026
mse0	11.5902	10.7458
rmse	3.40443	3.27808
mae	2.61826	2.48735
smape	12.0589	11.8858
m	392.000	78.0000
dfr	6.00000	6.00000
df	385.000	385.000
fStat	272.234	298.034
aic	-1022.45	-189.284
bic	-994.656	-172.787

Forward Selection:

(0.00000, 4.00000, 6.00000, 1.00000, 5.00000, 3.00000, 2.00000,  
-1359.19, -1125.97, -1031.56, -1029.34, -1027.03, -1025.01, -1022.45)

Backward Elimination:

(0.00000, 4.00000, 6.00000, 1.00000, 5.00000, 3.00000, 2.00000,  
-1359.19, -1125.97, -1031.56, -1029.34, -1027.03, -1025.01, -1022.45)

Stepwise Selection:

(0.00000, 4.00000, 6.00000, 1.00000, 5.00000, 3.00000,  
-1.00000, -1.00000, -1.00000, -1.00000, -1.00000)

### 2.2 House Prices

Forward Selection:

(0.00000, 1.00000, 4.00000, 5.00000, 2.00000, 3.00000, 6.00000,  
-13859.8, -11832.2, -11640.1, -11299.1, -10853.0, -10676.9, -10591.2)

Table 2: Statsmodels - Auto MPG Linear Regression

Metric	In-Sample	80-20 Split
rSq	0.8093	0.8223
rSqBar	0.8063	0.8075
sst	23818.9935	4032.2061
sse	4543.3470	716.3916
sde	3.4352	3.1543
mse0	11.8009	9.9499
rmse	3.4352	3.1543
mae	2.6183	2.2526
smape	12.0589	10.0104
m	392.0000	79.0000
dfr	6.0000	6.0000
df	385.0000	72.0000
fStat	272.2341	55.5419
aic	2086.9095	412.3698
bic	2114.7083	428.9560

Table 3: Scalation - Auto MPG Linear Regression CV

Name	num	min	max	mean	stdev	interval
rSq	5	0.788	0.823	0.798	0.014	0.018
rSqBar	5	0.785	0.820	0.795	0.014	0.018
sst	5	3962.818	5671.580	4700.481	620.767	770.935
sse	5	824.554	1176.435	950.494	142.696	177.215
sde	5	3.177	3.738	3.431	0.226	0.281
mse0	5	10.571	15.083	12.186	1.829	2.272
rmse	5	3.251	3.884	3.483	0.256	0.318
mae	5	2.487	2.850	2.689	0.151	0.188
smape	5	11.886	12.905	12.372	0.427	0.530
m	5	78.000	78.000	78.000	0.000	0.000
dfr	5	6.000	6.000	6.000	0.000	0.000
df	5	385.000	385.000	385.000	0.000	0.000
fStat	5	239.054	298.034	254.430	24.517	30.448
aic	5	-205.110	-188.647	-194.539	6.676	8.291
bic	5	-188.613	-172.150	-178.042	6.676	8.291

Backward Elimination:

(0.00000, 1.00000, 4.00000, 5.00000, 2.00000, 3.00000, 6.00000,  
-13859.8, -11832.2, -11640.1, -11299.1, -10853.0, -10676.9, -10591.2)

Stepwise Selection:

(0.00000, 1.00000, 4.00000, 5.00000, 2.00000, 3.00000,  
-1.00000, -1.00000, -1.00000, -1.00000, -1.00000)

## 2.3 Insurance Charges

Forward Selection:

(0.00000, 5.00000, 1.00000, 2.00000, 3.00000, 7.00000, 4.00000, 6.00000,  
-14475.6, -13826.7, -13616.7, -13548.9, -13541.0, -13537.9, -13535.8, -13533.8)

Backward Elimination:

(0.00000, 5.00000, 1.00000, 2.00000, 3.00000, 7.00000, 4.00000, 6.00000,  
-14475.6, -13826.7, -13616.7, -13548.9, -13541.0, -13537.9, -13535.8, -13533.8)

Stepwise Selection:

(0.00000, 5.00000, 1.00000, 2.00000, 3.00000, 7.00000, 4.00000,

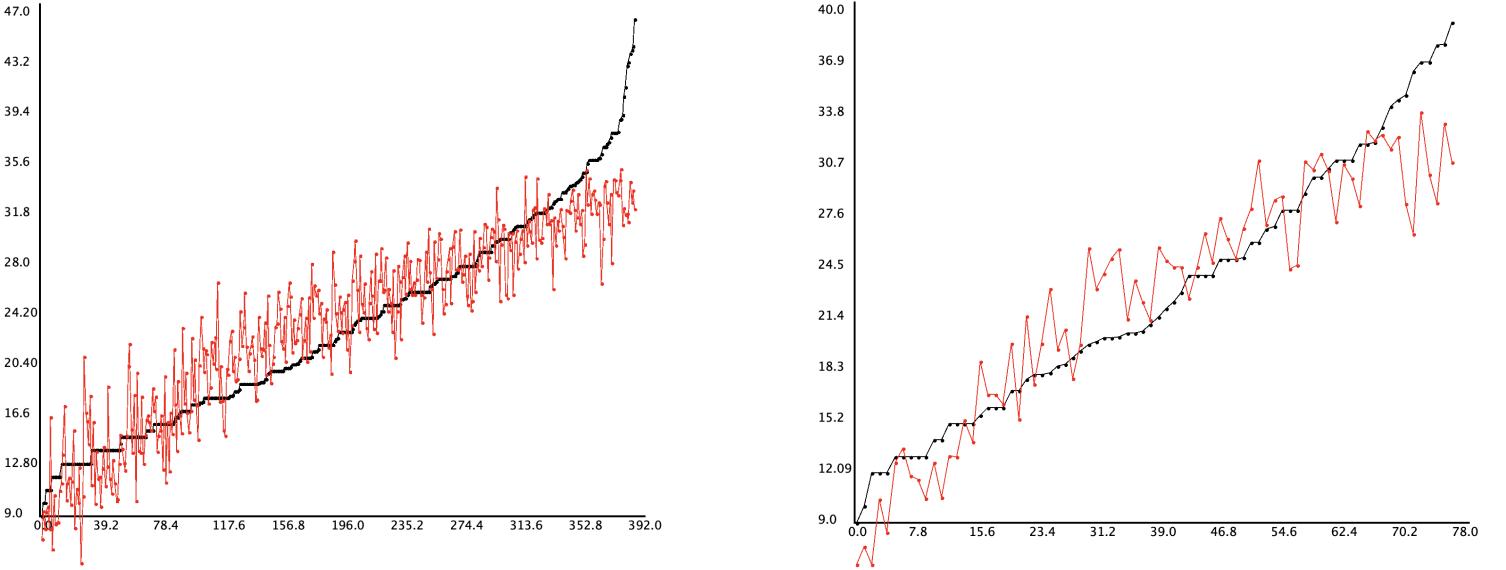


Figure 1: Scalation - Auto MPG Regression

Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 $yy$  black/actual vs.  $yp$  red/predicted

-1.00000, -1.00000, -1.00000, -1.00000, -1.00000, -1.00000, -1.00000)

### 3 Ridge Regression

#### 3.1 Auto MPG

#### 3.2 House Prices

#### 3.3 Insurance Charges

### Transformed Regression: Sqrt, Log1p, Box-Cox

In this section, we applied three different transformation techniques to the response variables (mpg, house price, and insurance charge) across the Auto MPG, Boston House Price, and Medical Cost datasets. The distributions for mpg, house price, and medical charge exhibit right-skewed patterns. Therefore, these transformations could significantly improved model accuracy. To measure and compare the quality of fit for the different transformation models, we extracted 15 metrics, including  $R^2$ , adjusted  $R^2$ , MSE, and MAE. Each model was evaluated using both in-sample validation and a validation set with an 80-20% split. Finally, we applied feature selection to identify the optimal set of features that best describe the response variable. For box-cox transformation we required a optimal lambda parameter which was 0.19, 0.85, and 0.04 for mpg, house price, and insurance cost respectively.

### Transformed Regression using Scala

We used TranRegression function form the modeling to perfrom Sqrt, Log1p, Box-Cox on the Auto MPG, Boston House Price, and Medical Cost datasets. The function takes predictor variables, a response variable, feature names, a factorization method, and transformation and inverse transformation methods as input, and fits the model based on the specified transformation. An example of the code used to fit the square-root (sqrt) transformation model is provided below.

```
f = ("sqrt", sq, "sqrt")
val mod = new modeling.TranRegression (ox, y, ox_fname,
                                         modeling.Regression.hp,
                                         f._1, f._2, f._3)
```

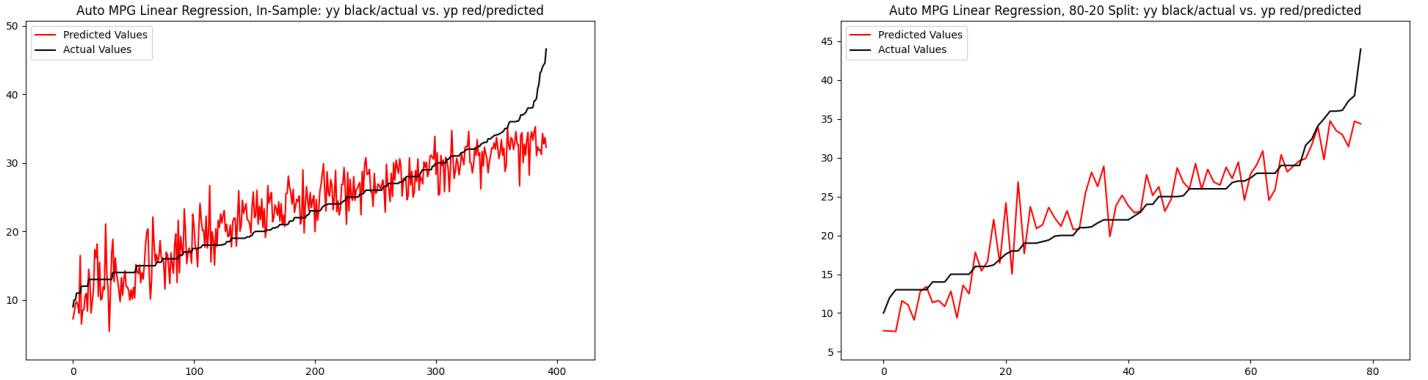


Figure 2: Statsmodels - Auto MPG Regression  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 $yy \text{ black/actual vs. } yp \text{ red/predicted}$

Table 4: Statsmodels - Auto MPG Linear Regression CV

Name	In-num folds	min	max	mean	stdev
rSq	5	0.8029	0.8223	0.8102	0.0067
rSqBar	5	0.7990	0.8188	0.8065	0.0068
sst	5	18025.6506	19750.2199	19045.3459	599.2336
sse	5	3202.8151	3813.1275	3617.5551	214.7409
sde	5	3.2300	3.5243	3.4334	0.1050
mse0	5	10.4326	12.4206	11.7994	0.7068
rmse	5	3.2300	3.5243	3.4334	0.1050
mae	5	2.5039	3.1904	2.6786	0.2601
smape	5	11.2379	14.1325	12.3805	0.9795
m	5	78.0000	79.0000	78.4000	0.4899
dfr	5	6.0000	6.0000	6.0000	0.0000
df	5	306.0000	307.0000	306.6000	0.4899
fStat	5	208.4306	236.8026	218.4653	9.8041
aic	5	1634.3246	1689.0924	1670.2523	18.7268
bic	5	1660.5704	1715.3381	1696.4892	18.7239

## Transformed Regression on the Auto MPG Dataset using Scala: In-Sample and Validation Results

Tables 1 and 2 present the quality of fit for the in-sample and validation (80-20% split) evaluations using the Auto MPG data. For the in-sample case, the models utilized all available data. The results indicate that the log1p and Box-Cox transformations outperform the sqrt transformation, which is further confirmed by the adjusted  $R^2$  values. The Mean Square Error (MSE) for the sqrt, log1p, and Box-Cox transformations reached 10.02, 9.17, and 9.40, respectively, with the log1p transformation achieving the lowest error. Similarly, in the validation accuracy results, where 80% of the data was randomly used for model training and 20% for testing, the log1p transformation achieved the highest  $R^2$  and adjusted  $R^2$  values, as well as the lowest MSE.

## Transformed Regression on the Boston House Price Dataset: In-Sample and Validation Results

For the house price prediction, the log1p transformed model performed significantly worse than the sqrt and Box-Cox transformations. The Box-Cox and sqrt models performed similarly, although Box-Cox achieved highest  $R^2$  and adjusted  $R^2$  of 0.998 in the in-sample evaluation. A similar trend was observed in the validation, where Box-Cox and sqrt outperformed the log1p transformation, with Box-Cox achieving the highest  $R^2$  and adjusted  $R^2$ .

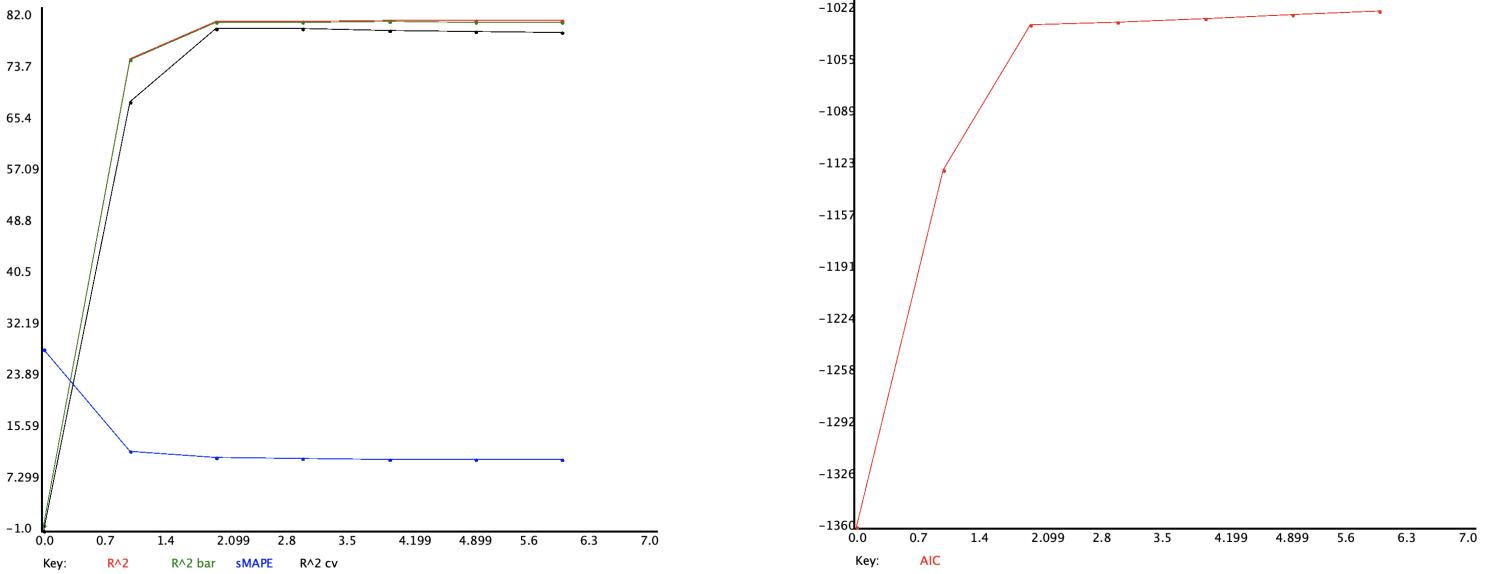


Figure 3: Scalation - Auto MPG Regression Forward Selection

Left:  $R^2$  vs.  $n$

Right: aic vs.  $n$

## Transformed Regression on the Medical Cost Dataset: In-Sample and Validation Results

In the Medical Cost dataset, the sqrt transformation performed significantly better than the log1p and Box-Cox transformations. The sqrt transformation achieved an  $R^2$  of 0.753 and an adjusted  $R^2$  of 0.751 in the in-sample evaluation, compared to 0.730 and 0.729, respectively, for the validation set. The adjusted  $R^2$  also shows a similar trend, where sqrt, log1p, and Box-Cox achieved 0.751, 0.520, and 0.574 for in-sample evaluation and 0.729, 0.569, and 0.607 for validation, respectively.

## Transformed Regression using statsmodels

Here, we used Python statsmodels library to reproduce all results generated by the TranRegression package.

### Transformed Regression with Sqrt, Log1p, and Box–Cox Transformations on the Auto MPG Dataset: In-Sample, Validation, Forward, and Backward Results

The following three tables present the results for the sqrt, log1p, and Box-Cox transformations, including in-sample evaluation, validation, and forward and backward feature selection. From these tables, we can see that, similar to the Scala results, the log1p and Box-Cox transformations perform similarly, with log1p being slightly better than Box-Cox. For the in-sample case, the sqrt, log1p, and Box-Cox transformations achieved  $R^2$  values of 0.835, 0.849, and 0.845, respectively, with adjusted  $R^2$  values of 0.833, 0.847, and 0.843. Similarly, for the validation case, the  $R^2$  values were 0.822, 0.832, and 0.830, with adjusted  $R^2$  values of 0.807, 0.818, and 0.816.

### Transformed Regression with Sqrt, Log1p, and Box–Cox Transformations on the Boston House Price Dataset: In-Sample, Validation, Forward, and Backward Results

For the house price prediction dataset, the ScalaTion and statsmodels summaries are consistent, as the Box–Cox transformation was identified as the best-performing model under both approaches.

### Transformed Regression with Sqrt, Log1p, and Box–Cox Transformations on the Medical Cost Dataset: In-Sample, Validation, Forward, and Backward Results

Finally, for the medical cost dataset, the sqrt transformation explained the insurance costs better than the other methods, despite the relatively lower  $R^2$  values. For the in-sample case, the  $R^2$  values were 0.753, 0.523, and 0.576 for the sqrt, log1p, and Box-Cox transformations, respectively, with adjusted  $R^2$  values of 0.751, 0.520, and 0.574. These results were verified by

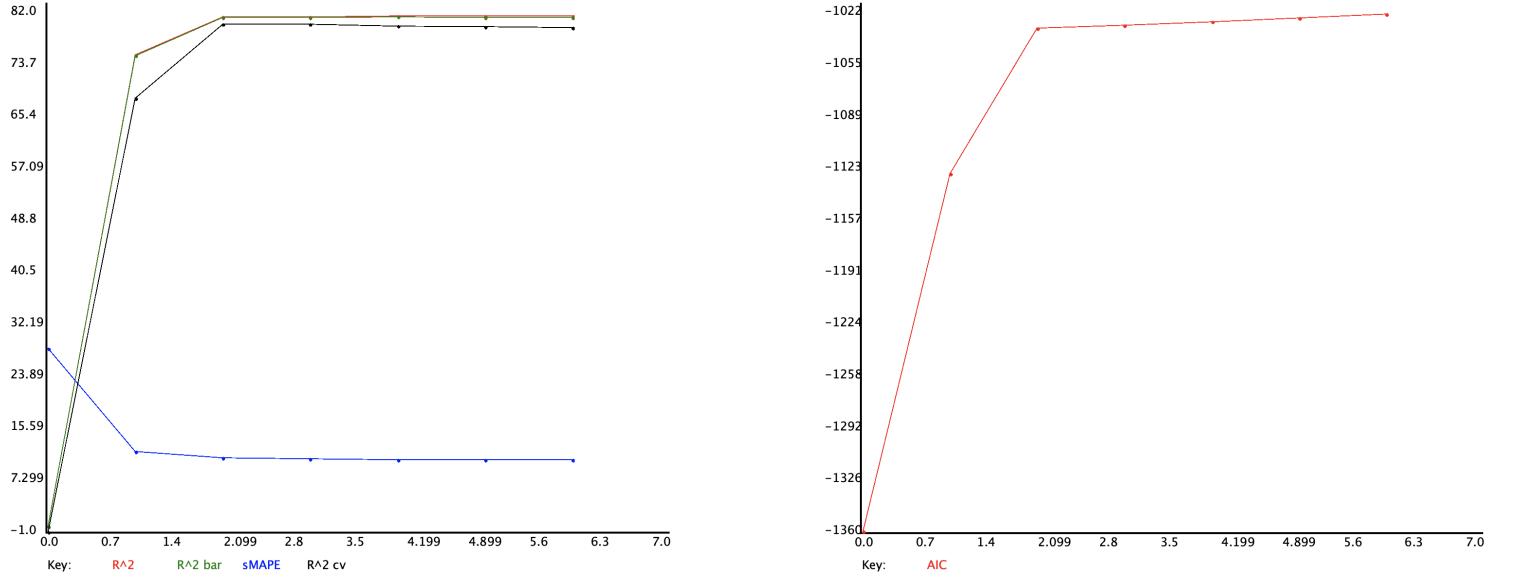


Figure 4: Scalation - Auto MPG Regression Backward Elimination

Left:  $R^2$  vs.  $n$

Right: aic vs.  $n$

the validation set, where the sqrt transformation achieved an  $R^2$  of 0.706 and an adjusted  $R^2$  of 0.697, outperforming the log1p ( $R^2$  of 0.505, Adjusted  $R^2$  of 0.490) and Box-Cox (( $R^2$  of 0.546, Adjusted  $R^2$  of 0.532) models.

## 4 Comparing Different Models on the Same Data Set

### 4.1 Auto MPG

### 4.2 House Prices

### 4.3 Insurance Charges

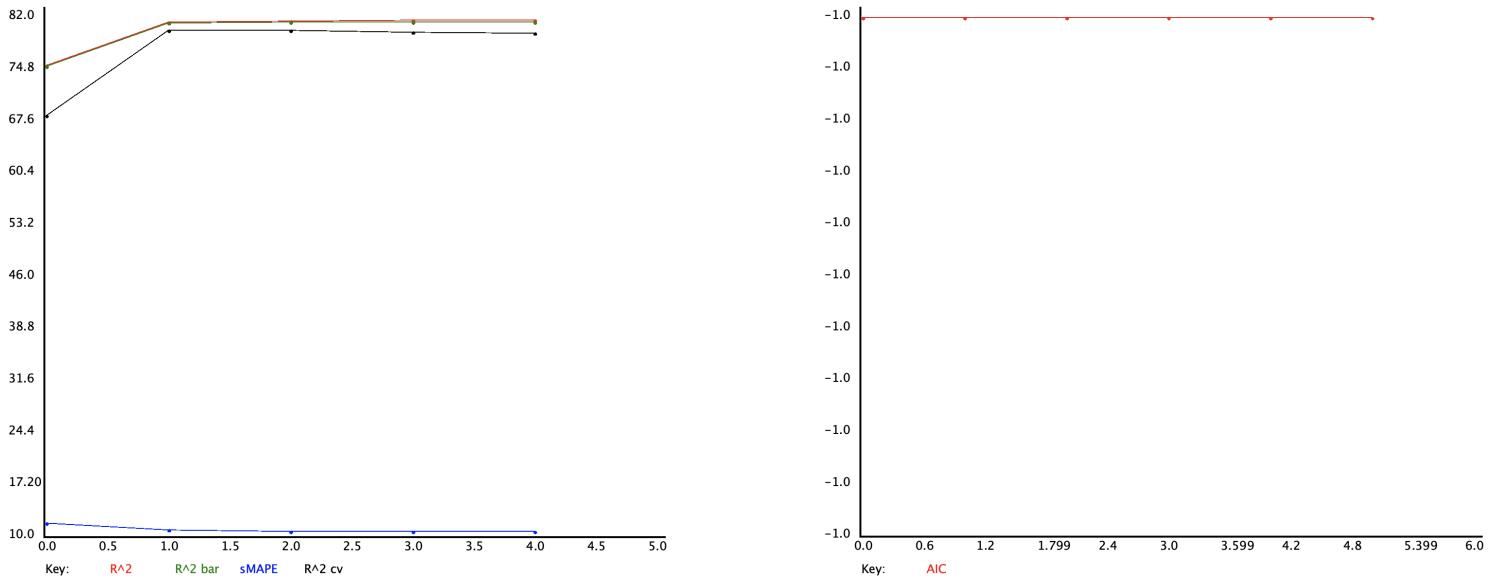


Figure 5: Scalation - Auto MPG Regression Stepwise Selection  
 Left:  $R^2$  vs.  $n$   
 Right: aic vs.  $n$

Table 5: Scalation - House Price Linear Regression

Metric	In-Sample	80-20 Split
rSq	0.998516	0.998649
rSqBar	0.998507	0.998641
sst	6.42325e+13	1.33700e+13
sse	9.53030e+10	1.80638e+10
sde	9767.21	9501.56
mse0	9.53030e+07	9.03190e+07
rmse	9762.32	9503.63
mae	7747.66	7484.48
smape	1.57791	1.60847
m	1000.00	200.000
dfr	6.00000	6.00000
df	993.000	993.000
fStat	111378	122330
aic	-10591.2	-2101.67
bic	-10556.9	-2078.59

Table 6: Statsmodels - House Price Linear Regression

Metric	In-Sample	80-20 Split
rSq	0.9985	0.9985
rSqBar	0.9985	0.9984
sst	64232463468052.5469	12891771417242.6445
sse	95249090298.3967	19682384464.0853
sde	9798.8381	10124.8417
mse0	96017228.1234	102512419.0838
rmse	9798.8381	10124.8417
mae	7740.4301	7898.0257
smape	1.5774	1.6060
m	1000.0000	200.0000
dfr	7.0000	7.0000
df	992.0000	192.0000
fStat	95425.1583	17938.0204
aic	21225.8831	4264.5099
bic	21265.1451	4290.8965

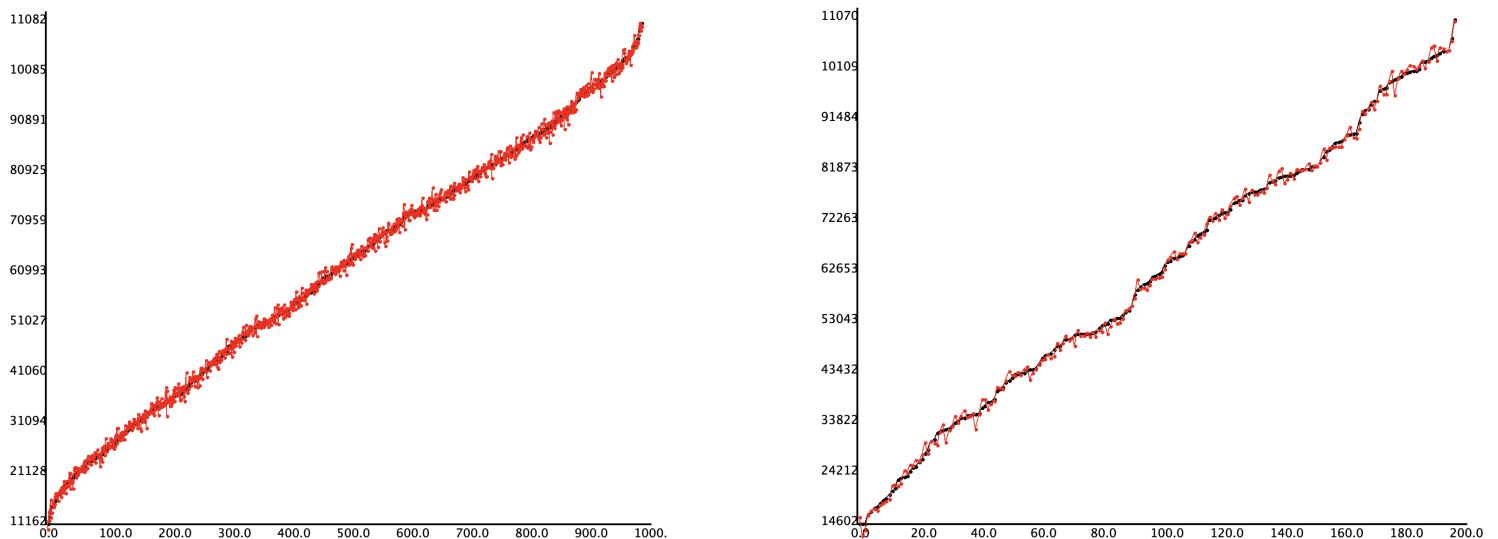


Figure 6: Scalation - House Price Regression

Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

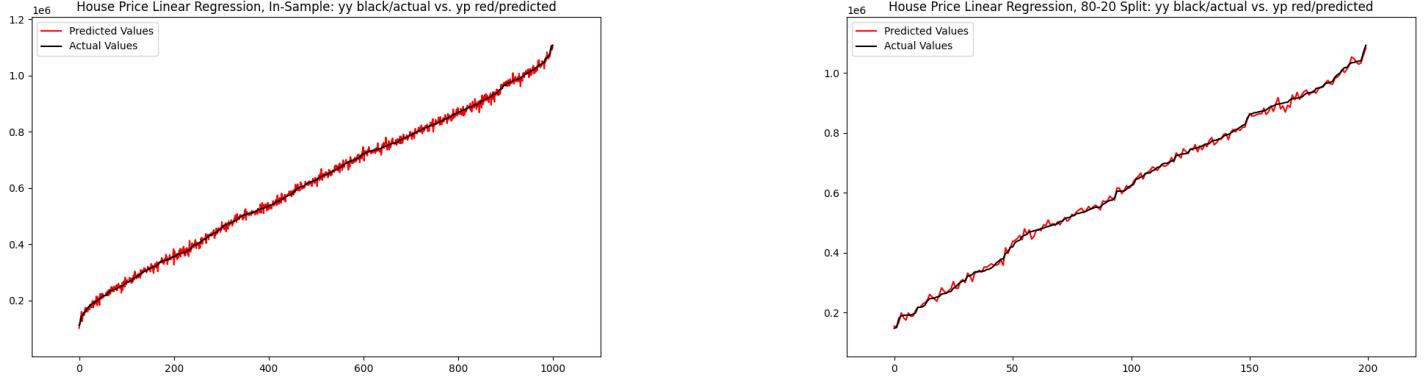


Figure 7: Statsmodels - House Price Regression

Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 $yy \text{ black/actual vs. } yp \text{ red/predicted}$

Table 7: Scalation - House Price Linear Regression CV

Name	num folds	min	max	mean	stdev	interval
rSq	5	0.998	0.999	0.998	0.000	0.000
rSqBar	5	0.998	0.999	0.998	0.000	0.000
sst	5	11805100457351.200	13369989427686.710	12829460198984.865	606159535374.424	752793908917.133
sse	5	17790794374.705	22663619750.593	19394433560.564	2117948592.845	2630295668.134
sde	5	9365.942	10608.530	9818.348	533.068	662.021
mse0	5	88953971.874	113318098.753	96972167.803	10589742.964	13151478.341
rmse	5	9431.541	10645.097	9836.093	528.486	656.330
mae	5	7418.536	8609.308	7817.189	534.033	663.219
smape	5	1.408	1.804	1.592	0.141	0.176
m	5	200.000	200.000	200.000	0.000	0.000
dfr	5	6.000	6.000	6.000	0.000	0.000
df	5	993.000	993.000	993.000	0.000	0.000
fStat	5	96001.466	122329.999	110142.703	10843.190	13466.236
aic	5	-2127.278	-2100.155	-2109.082	11.789	14.641
bic	5	-2104.190	-2077.067	-2085.993	11.789	14.641

Table 8: Statsmodels - House Price Linear Regression CV

Name	In-num folds	min	max	mean	stdev
rSq	5	0.9985	0.9985	0.9985	0.0000
rSqBar	5	0.9985	0.9985	0.9985	0.0000
sst	5	50658229578994.8281	51833647639647.2656	51384861663931.4844	395867464990.6846
sse	5	74379785928.0215	78081632373.2809	76098283299.9349	1285733275.0049
sde	5	9690.9169	9929.1450	9801.8796	82.7530
mse0	5	93913871.1212	98587919.6632	96083691.0353	1623400.5998
rmse	5	9690.9169	9929.1450	9801.8796	82.7530
mae	5	7516.9137	8174.5836	7794.6342	224.6310
smape	5	1.5104	1.6620	1.5879	0.0540
m	5	200.0000	200.0000	200.0000	0.0000
dfr	5	7.0000	7.0000	7.0000	0.0000
df	5	792.0000	792.0000	792.0000	0.0000
fStat	5	74563.5320	77254.5038	76300.5459	963.0215
aic	5	16964.5723	17003.4288	16982.7315	13.4995
bic	5	17002.0492	17040.9057	17020.2084	13.4995

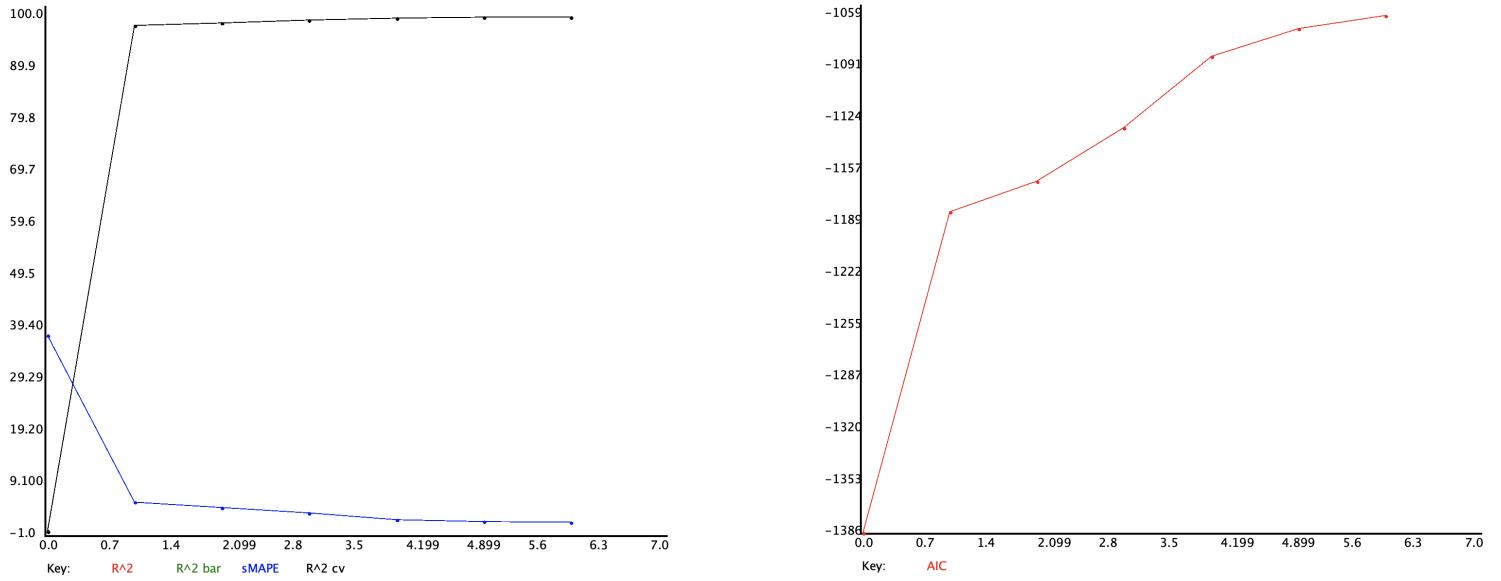


Figure 8: Scalation - House Prices Regression Forward Selection

Left:  $R^2$  vs.  $n$

Right: aic vs.  $n$

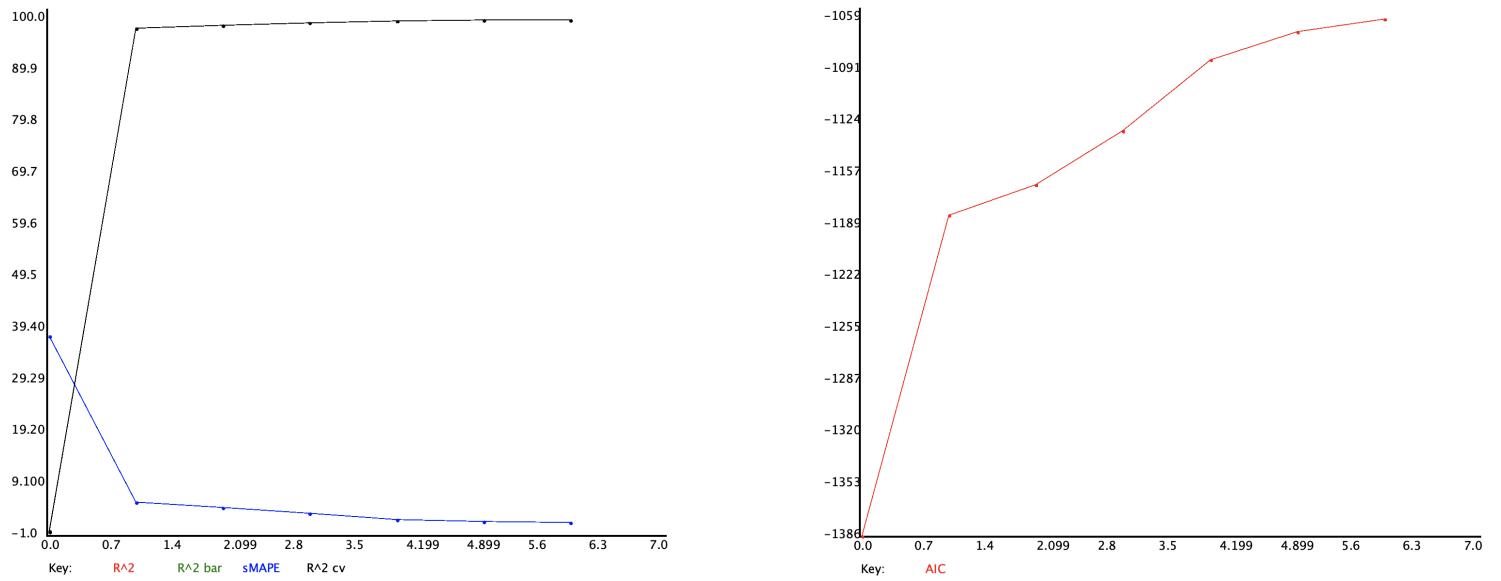


Figure 9: Scalation - House Prices Regression Backward Elimination

Left:  $R^2$  vs.  $n$   
Right: aic vs.  $n$

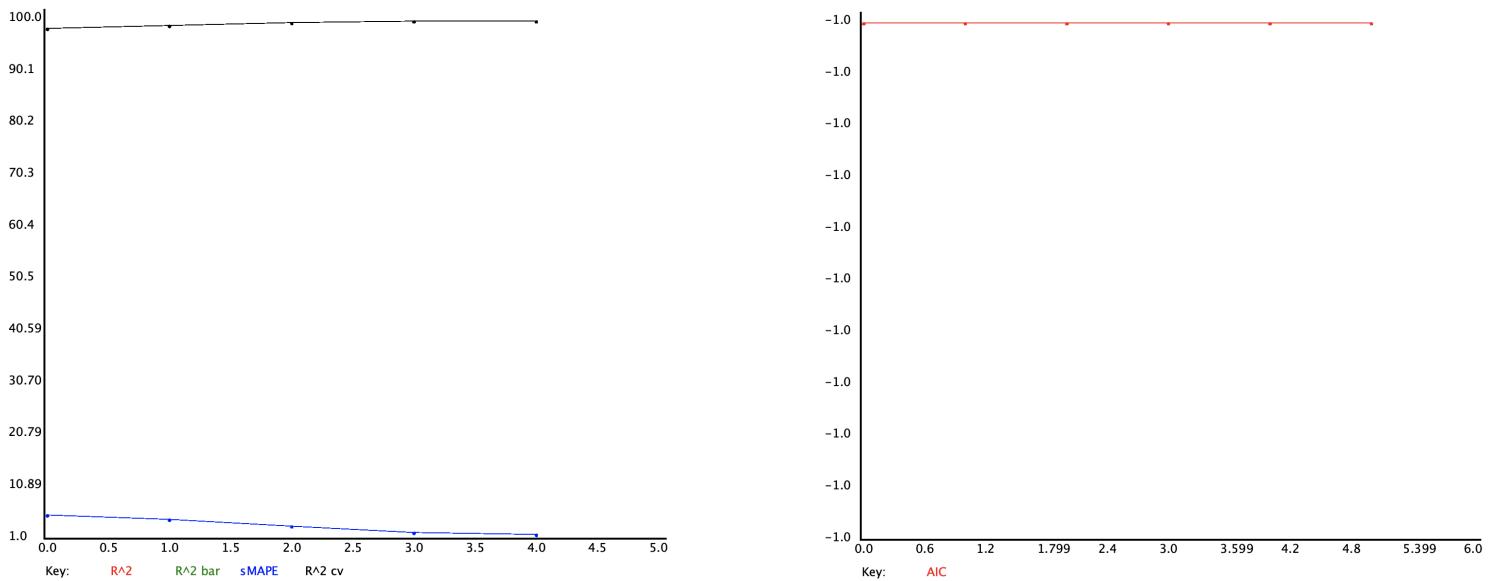


Figure 10: Scalation - House Prices Regression Stepwise Selection

Left:  $R^2$  vs.  $n$   
Right: aic vs.  $n$

Table 9: Scalation - Insurance Charges Linear Regression

Metric	In-Sample	80-20 Split
rSq	0.750157	0.720005
rSqBar	0.748842	0.718531
sst	1.96074e+11	4.06432e+10
sse	4.89878e+10	1.13799e+10
sde	6053.11	6540.46
mse0	3.66127e+07	4.26213e+07
rmse	6050.84	6528.50
mae	4179.54	4430.66
smape	37.9722	40.0602
m	1338.00	267.000
dfr	7.00000	7.00000
df	1330.00	1330.00
fStat	570.477	488.584
aic	-13533.8	-2708.17
bic	-13492.2	-2679.47

Table 10: Statsmodels - Insurance Charges Linear Regression

Metric	In-Sample	80-20 Split
rSq	0.7509	0.7898
rSqBar	0.7494	0.7833
sst	196074221568.3671	41606660039.7953
sse	48839532843.9219	8744084438.2470
sde	6062.1023	5810.4168
mse0	36749084.1564	33760943.7770
rmse	6062.1023	5810.4168
mae	4170.8869	3959.0670
smape	37.8059	37.7432
m	1338.0000	268.0000
dfr	8.0000	8.0000
df	1329.0000	259.0000
fStat	500.8107	121.6738
aic	27113.5058	5415.1269
bic	27160.2962	5447.4458

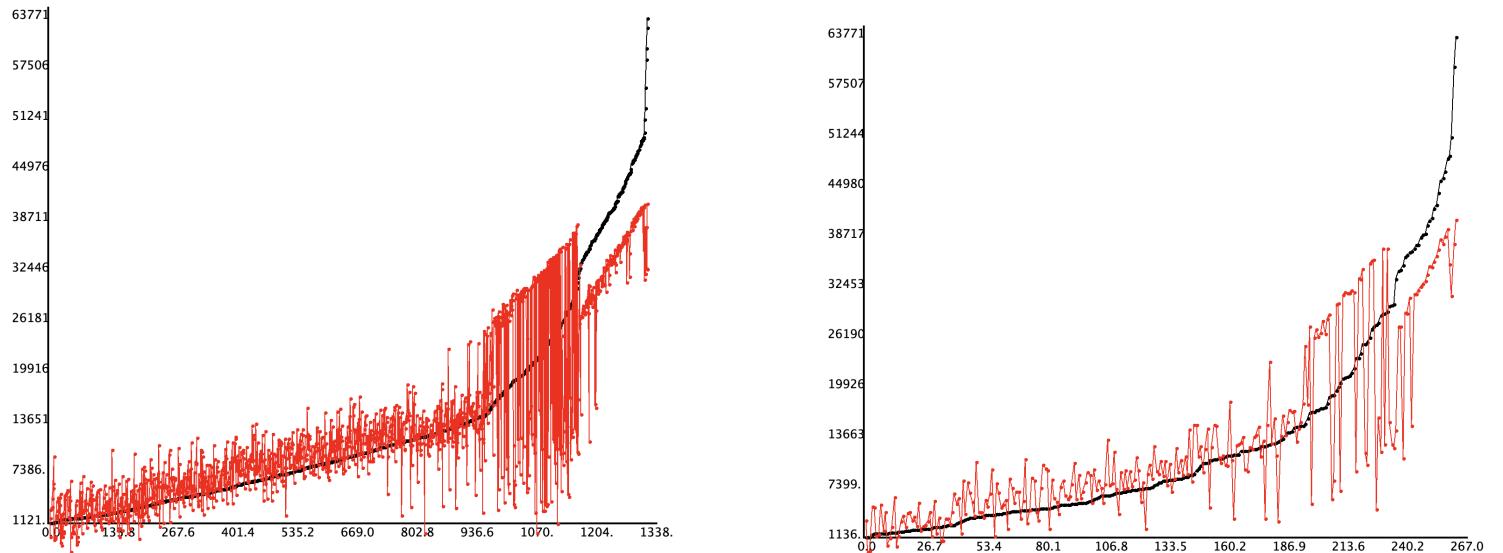


Figure 11: Scalation - Insurance Charges Regression  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

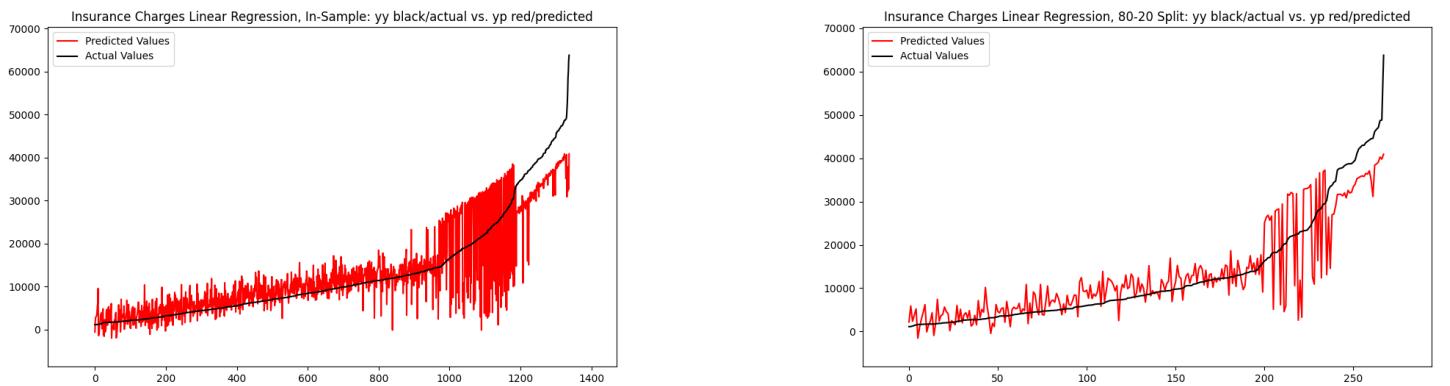


Figure 12: Statsmodels - Insurance Charges Regression  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

Table 11: Scalation - Insurance Charges Linear Regression CV

Name	num	min	max	mean	stdev	interval
rSq	5	0.701	0.814	0.743	0.046	0.057
rSqBar	5	0.699	0.813	0.742	0.046	0.057
sst	5	31902777173.848	43430486230.657	38949749086.422	4343029725.651	5393640012.108
sse	5	7539480548.420	11454966761.392	9918152392.401	1670500799.560	2074607019.047
sde	5	5320.957	6562.018	6084.654	526.821	654.263
mse0	5	28237754.863	42902497.234	37146638.174	6256557.302	7770063.742
rmse	5	5313.921	6550.000	6076.688	525.006	652.009
mae	5	3810.754	4511.498	4216.703	292.688	363.491
smape	5	36.128	40.060	38.143	1.833	2.277
m	5	267.000	267.000	267.000	0.000	0.000
dfr	5	7.000	7.000	7.000	0.000	0.000
df	5	1330.000	1330.000	1330.000	0.000	0.000
fStat	5	444.882	830.519	573.015	157.534	195.643
aic	5	-2709.047	-2663.110	-2691.017	19.599	24.340
bic	5	-2680.349	-2634.412	-2662.319	19.599	24.340

Table 12: Statsmodels - Insurance Charges Linear Regression CV

Name	In-num folds	min	max	mean	stdev
rSq	5	0.7372	0.7714	0.7509	0.0118
rSqBar	5	0.7352	0.7697	0.7490	0.0119
sst	5	152000096476.0009	165778288756.8451	156844205919.5975	4829672803.3539
sse	5	37889563873.8970	39941665228.4872	39013173504.6473	798799753.1118
sde	5	5973.0692	6135.5767	6062.3950	63.2693
mse0	5	35677555.4368	37645301.8176	36756635.6629	766907.3866
rmse	5	5973.0692	6135.5767	6062.3950	63.2693
mae	5	4054.1099	4427.9335	4203.4121	129.0554
smape	5	35.6194	40.0220	38.1279	1.5723
m	5	267.0000	268.0000	267.6000	0.4899
dfr	5	8.0000	8.0000	8.0000	0.0000
df	5	1061.0000	1062.0000	1061.4000	0.4899
fStat	5	372.0864	448.0714	401.2199	26.3875
aic	5	21673.0530	21710.2694	21692.5689	14.8342
bic	5	21717.8401	21755.0482	21737.3510	14.8317

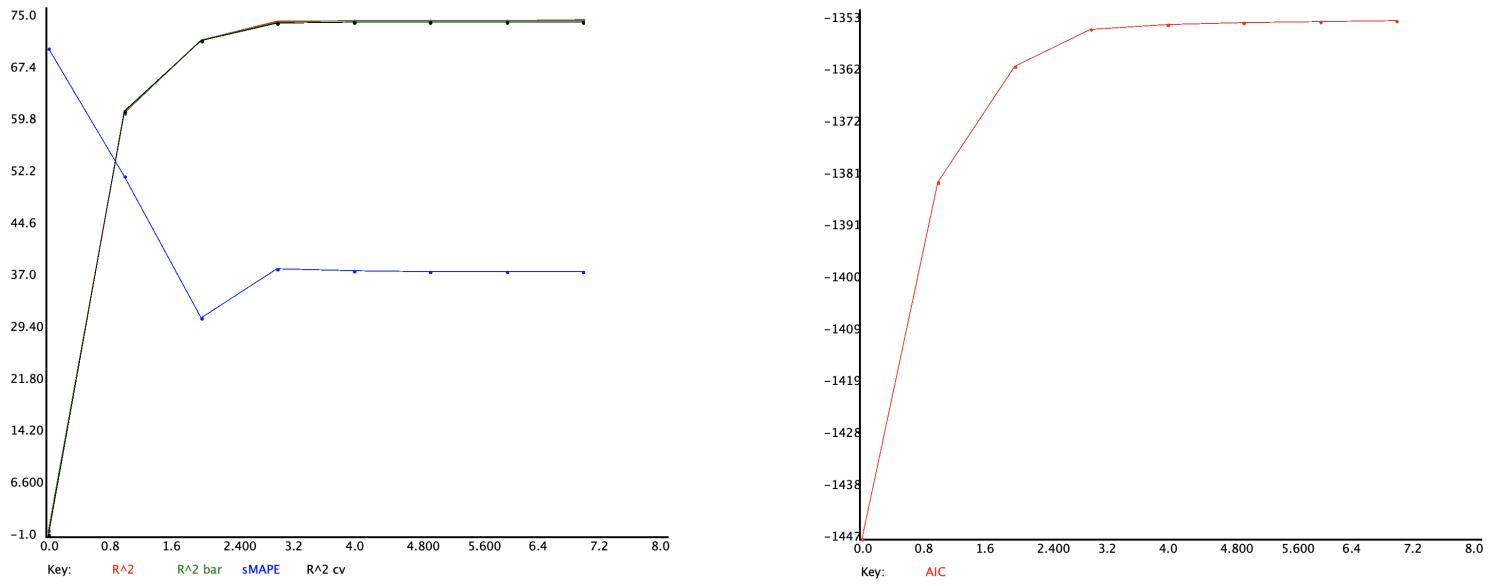


Figure 13: Scalation - Insurance Charges Regression Forward Selection

Left:  $R^2$  vs.  $n$   
Right: aic vs.  $n$

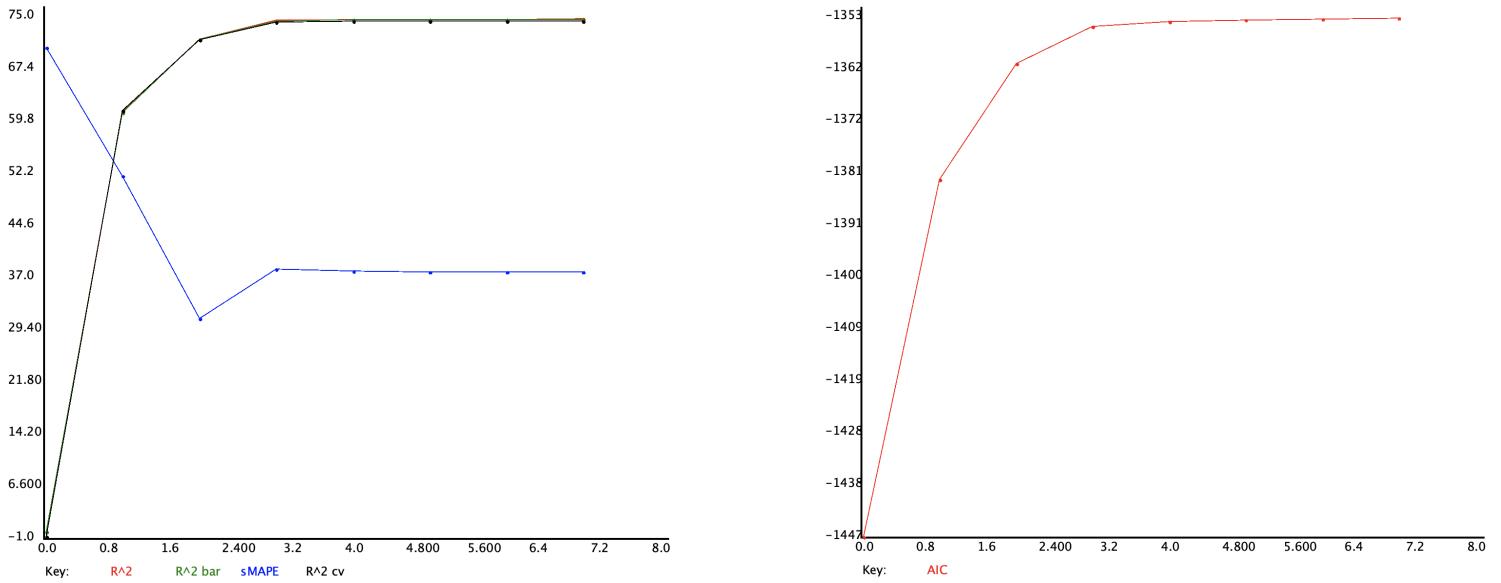


Figure 14: Scalation - Insurance Charges Regression Backward Elimination

Left:  $R^2$  vs.  $n$   
Right: aic vs.  $n$

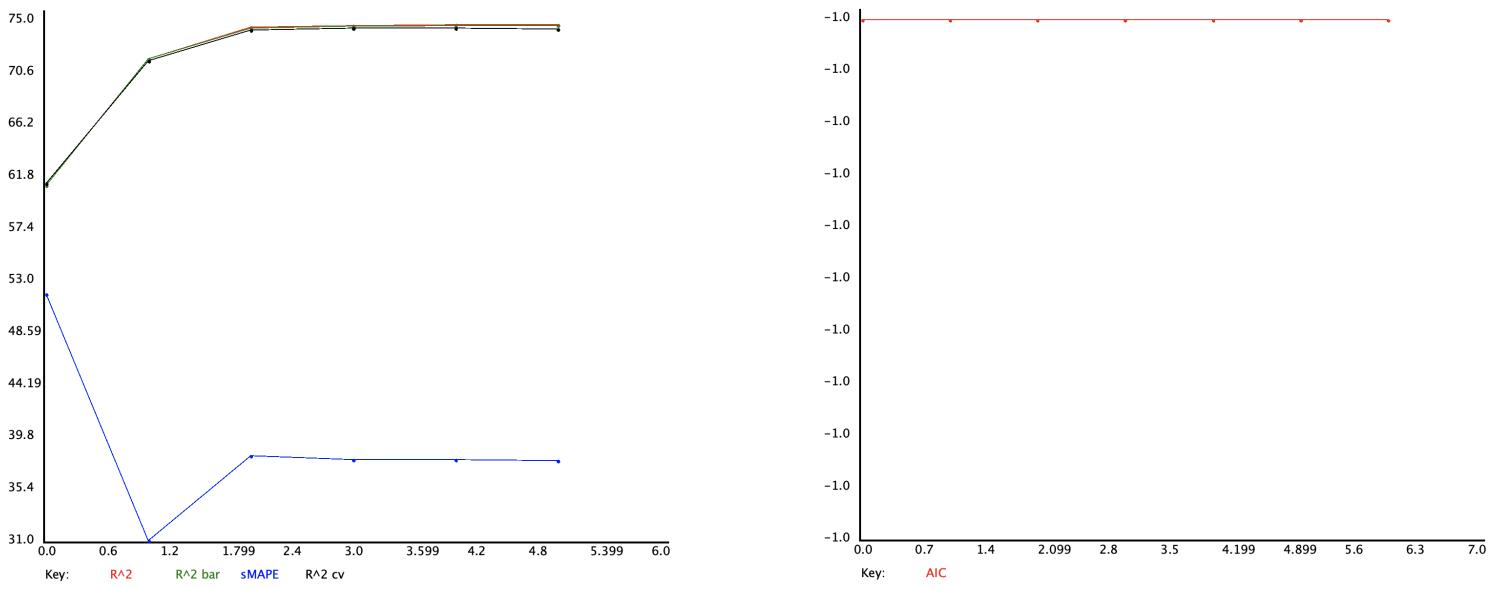


Figure 15: Scalation - Insurance Charges Regression Stepwise Selection  
 Left:  $R^2$  vs.  $n$   
 Right: aic vs.  $n$

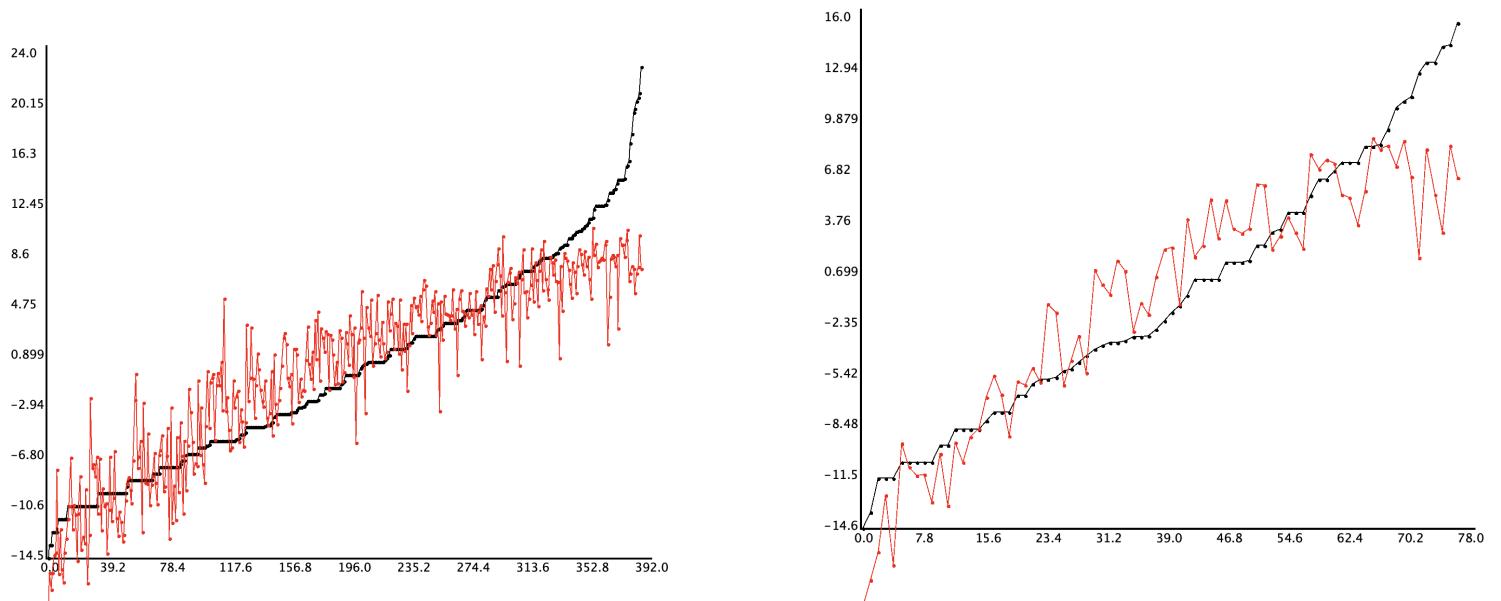


Figure 16: Scalation - Auto MPG Ridge  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

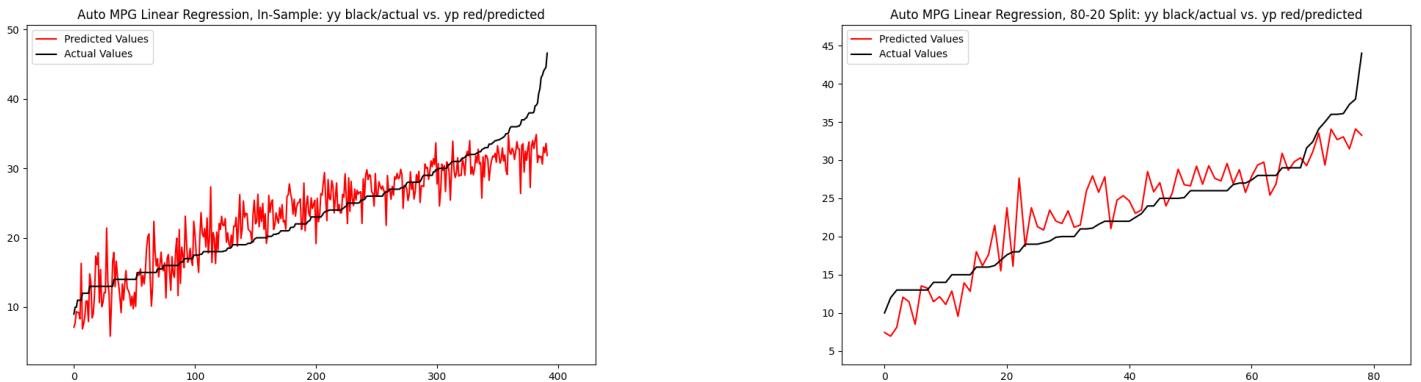


Figure 17: Statsmodels - Auto MPG Ridge

Left: In Sample Predictions

Right: 80-20 Out of Sample Predictions  
yy black/actual vs. yp red/predicted

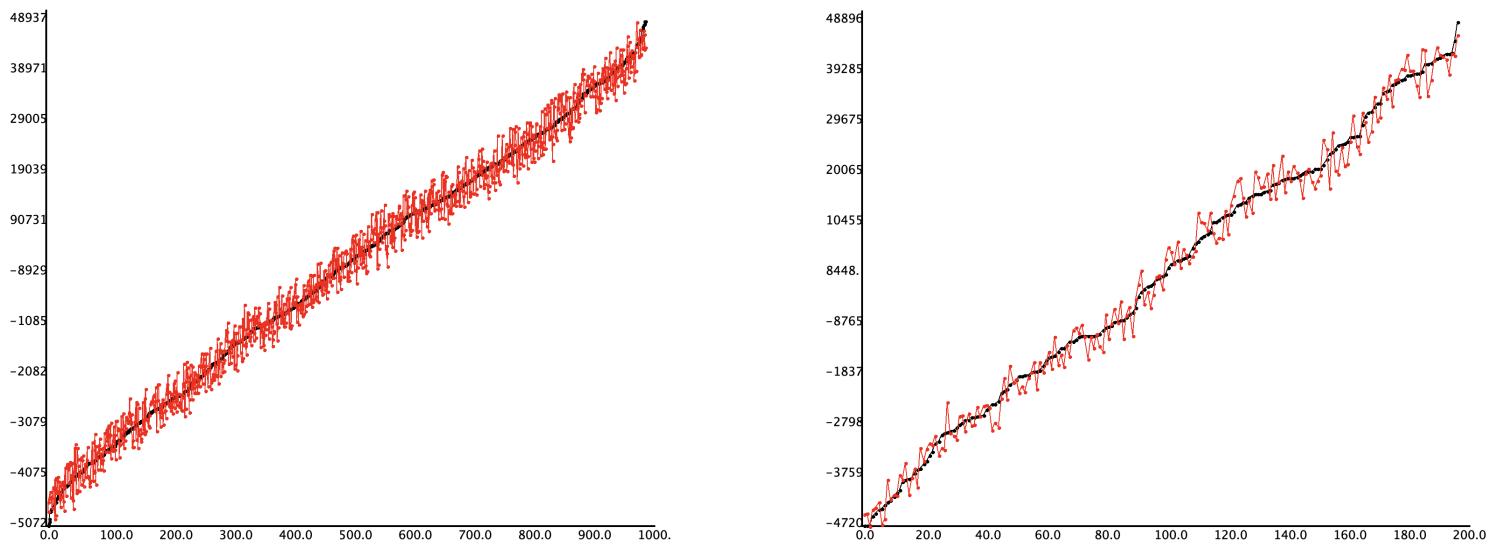


Figure 18: Scalation - House Price Ridge

Left: In Sample Predictions

Right: 80-20 Out-of-Sample Predictions  
yy black/actual vs. yp red/predicted

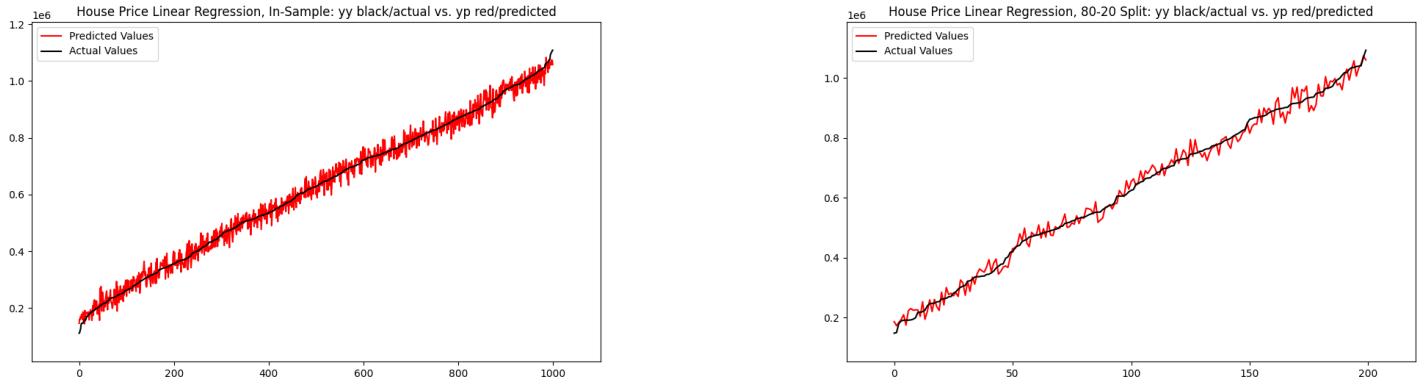


Figure 19: Statsmodels - House Price Ridge

Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 $yy \text{ black/actual vs. } yp \text{ red/predicted}$

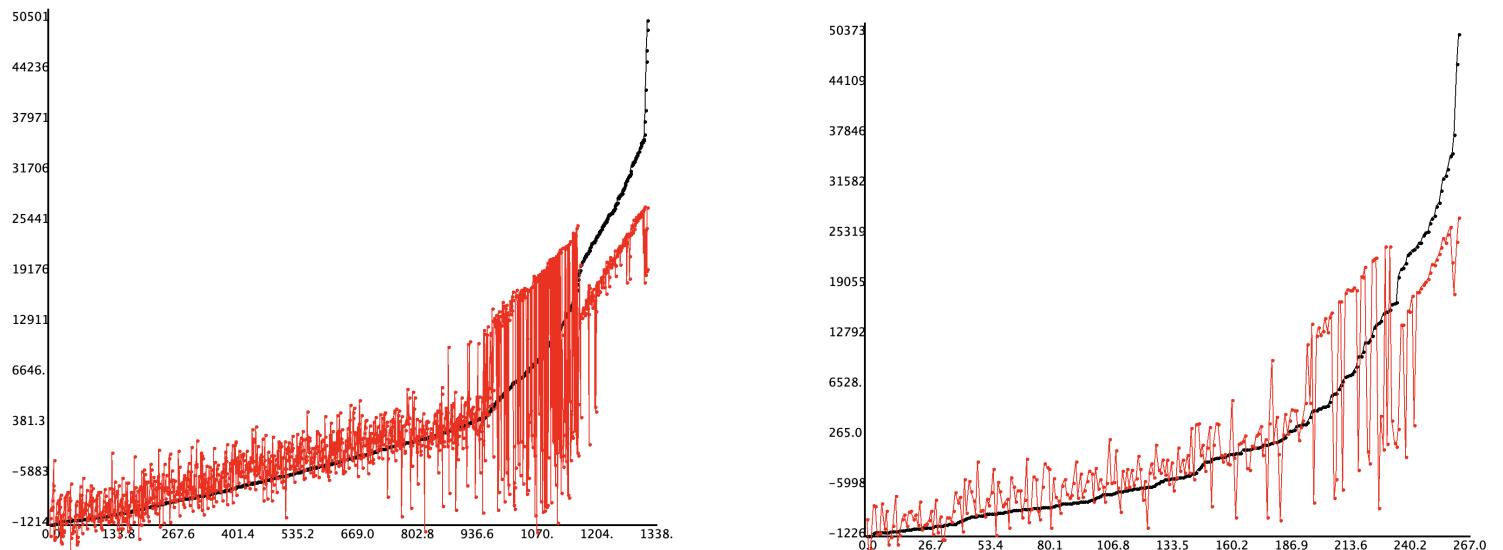


Figure 20: Scalation - Insurance Charges Ridge

Left: In Sample Predictions  
 Right: 80-20 Out-of-Sample Predictions  
 $yy \text{ black/actual vs. } yp \text{ red/predicted}$

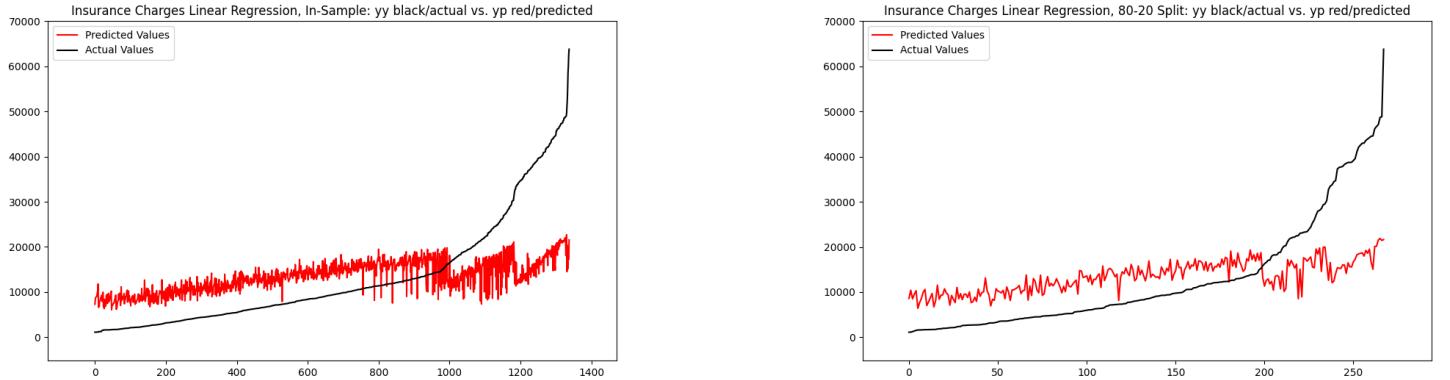


Figure 21: Statsmodels - Insurance Charges Ridge  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 $yy \text{ black/actual vs. } yp \text{ red/predicted}$

Table 13: Auto MPG (In-Sample): Sqrt, Log1p, and Box–Cox( $\lambda = 0.19$ )

Metric	sqrt	log1p	box-cox( $\lambda=0.19$ )
rSq	0.835138	0.849102	0.845366
rSqBar	0.832569	0.846751	0.842956
sst	23819.0	23819.0	23819.0
sse	3926.85	3594.23	3683.22
sde	3.16757	3.02514	3.06455
mse0	10.0175	9.16895	9.39596
rmse	3.16504	3.02803	3.06528
mae	2.34675	2.18422	2.22945
smape	10.2046	9.32001	9.54480
m	392.000	392.000	392.000
dfr	6.00000	6.00000	6.00000
df	385.000	385.000	385.000
fStat	325.048	361.067	350.793
aic	-993.873	-976.527	-981.320
bic	-966.074	-948.728	-953.521

Table 14: Auto MPG (Validation): Sqrt, Log1p, and Box–Cox( $\lambda = 0.19$ )

Metric	sqrt	log1p	box-cox( $\lambda=0.19$ )
rSq	0.846480	0.852864	0.851819
rSqBar	0.844088	0.850571	0.849510
sst	4731.23	4731.23	4731.23
sse	726.337	696.133	701.080
sde	3.05724	2.97969	2.99552
mse0	9.31202	8.92478	8.98820
rmse	3.05156	2.98744	2.99803
mae	2.11846	1.98665	2.01582
smape	9.00068	8.28831	8.41880
m	78.0000	78.0000	78.0000
dfr	6.00000	6.00000	6.00000
df	385.000	385.000	385.000
fStat	353.804	371.939	368.862
aic	-183.700	-182.048	-182.322
bic	-167.203	-165.551	-165.825

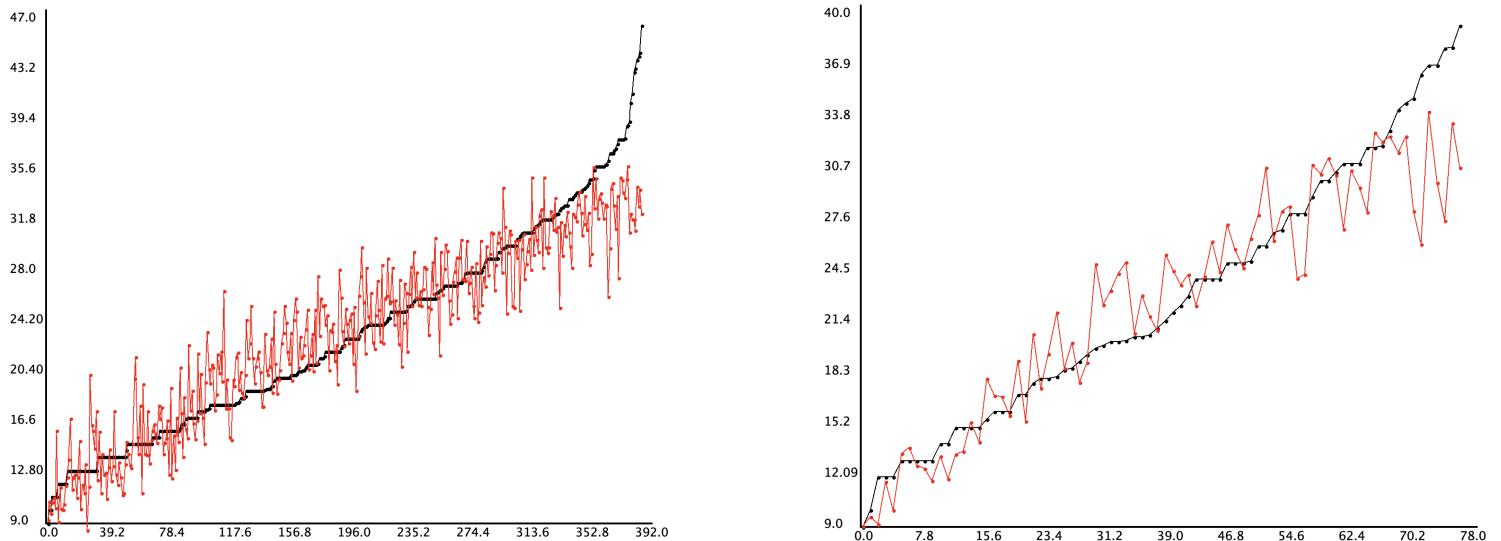


Figure 22: Scalation - Auto MPG Sqrt

Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

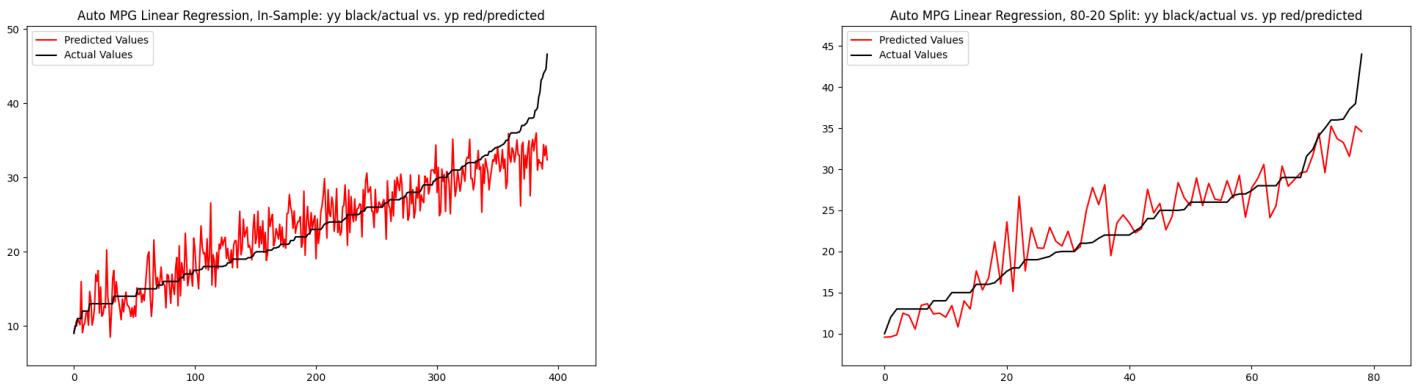


Figure 23: Statsmodels - Auto MPG Sqrt

Left: In Sample Predictions  
Right: 80-20 Out of Sample Predictions  
yy black/actual vs. yp red/predicted

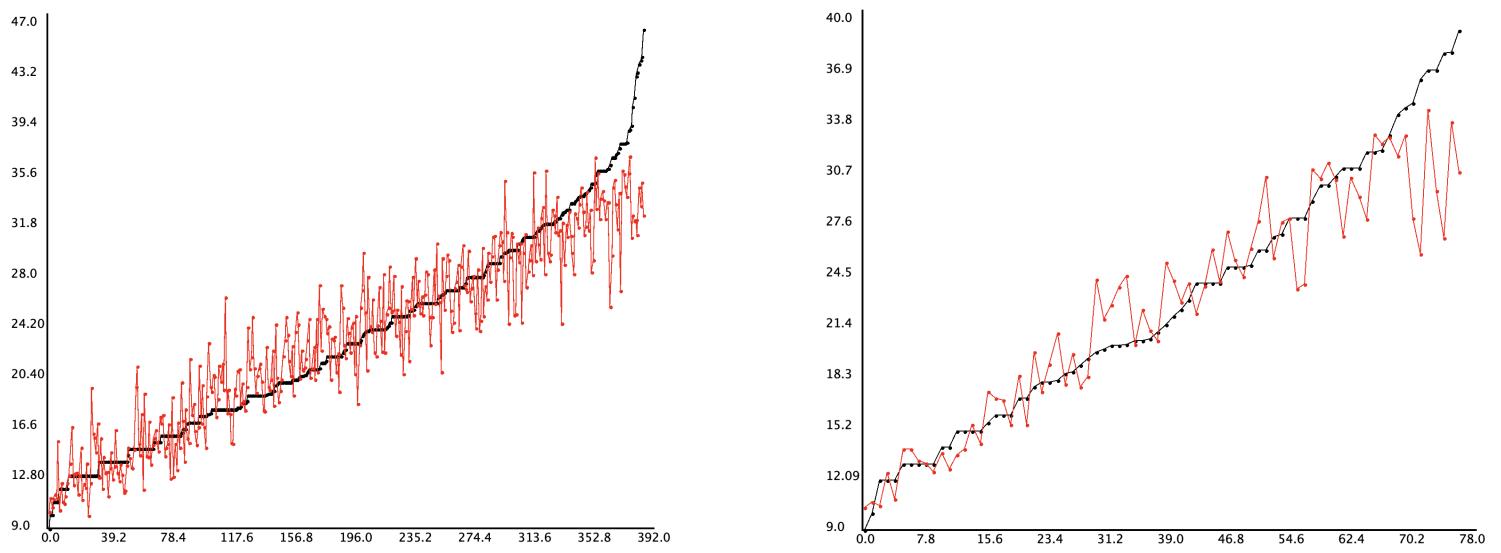


Figure 24: Scalation - Auto MPG log1p

Left: In Sample Predictions  
Right: 80-20 Out-of-Sample Predictions  
yy black/actual vs. yp red/predicted

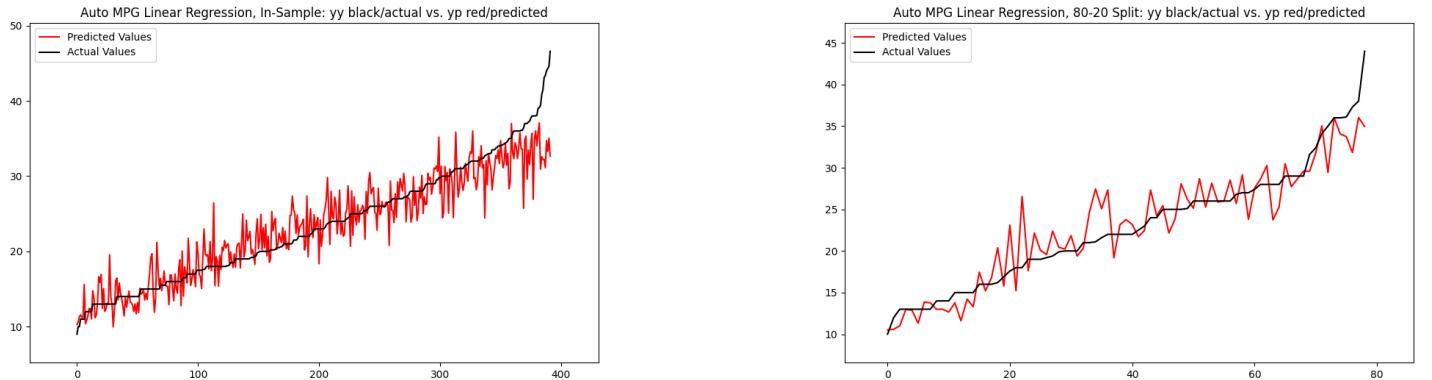


Figure 25: Statsmodels - Auto MPG log1p

Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 $yy$  black/actual vs.  $yp$  red/predicted

Table 15: House Price (In-Sample): Sqrt, Log1p, and Box–Cox( $\lambda = 0.85$ )

Metric	sqrt	log1p	box-cox( $\lambda=0.85$ )
rSq	0.986164	0.922376	0.997549
rSqBar	0.986067	0.921828	0.997531
sst	6.42325e+13	6.42325e+13	6.42325e+13
sse	8.88711e+11	4.98598e+12	1.57458e+11
sde	29823.1	70573.3	12554.0
mse0	8.88711e+08	4.98598e+09	1.57458e+08
rmse	29811.3	70611.5	12548.2
mae	24296.5	52989.3	10078.9
smape	4.85788	9.21218	2.12365
m	1000.00	1000.00	1000.00
dfr	7.00000	7.00000	7.00000
df	992.000	992.000	992.000
fStat	10100.8	1683.94	57668.5
aic	-11705.6	-12567.9	-10840.3
bic	-11666.3	-12528.6	-10801.0

Table 16: House Price (Validation): Sqrt, Log1p, and Box–Cox( $\lambda = 0.85$ )

Metric	sqrt	log1p	box-cox( $\lambda=0.85$ )
rSq	0.984017	0.906459	0.997398
rSqBar	0.983904	0.905799	0.997380
sst	1.33700e+13	1.33700e+13	1.33700e+13
sse	2.13694e+11	1.25064e+12	3.47839e+10
sde	32755.5	78812.3	13217.8
mse0	1.06847e+09	6.25321e+09	1.73920e+08
rmse	32687.5	79077.2	13187.9
mae	26140.1	58024.9	10655.9
smape	5.18595	9.88665	2.22264
m	200.000	200.000	200.000
dfr	7.00000	7.00000	7.00000
df	992.000	992.000	992.000
fStat	8724.77	1373.28	54329.4
aic	-2346.74	-2523.43	-2165.20
bic	-2320.35	-2497.04	-2138.81

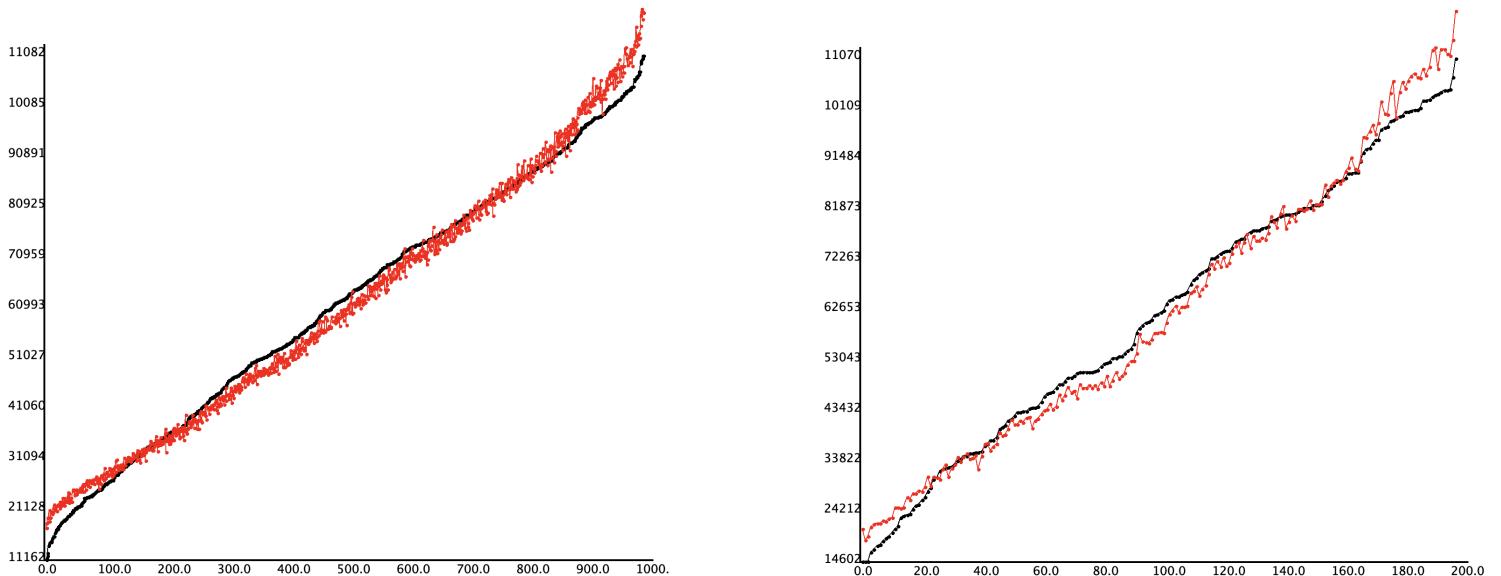


Figure 26: Scalation - House Price Sqrt

Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

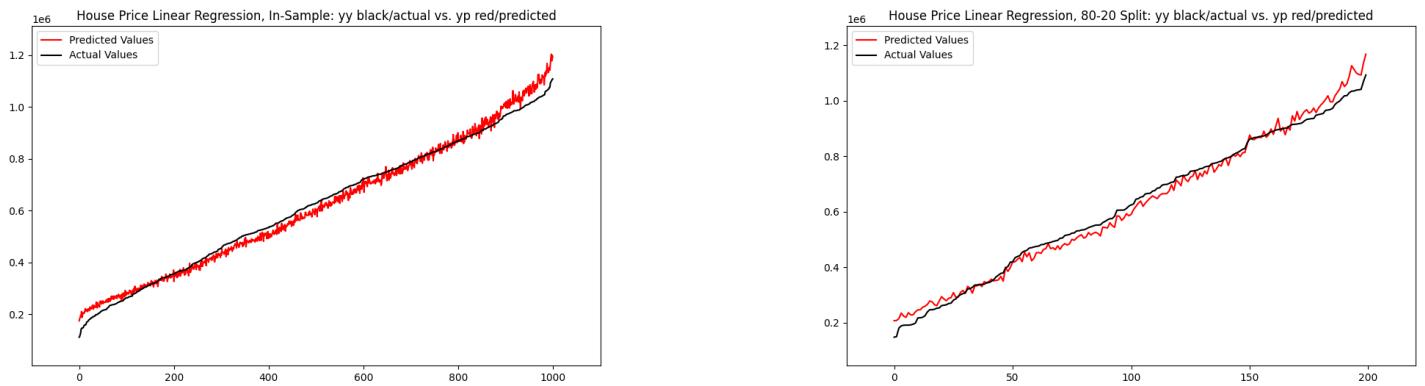


Figure 27: Statsmodels - House Price Sqrt

Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 $yy \text{ black/actual vs. } yp \text{ red/predicted}$

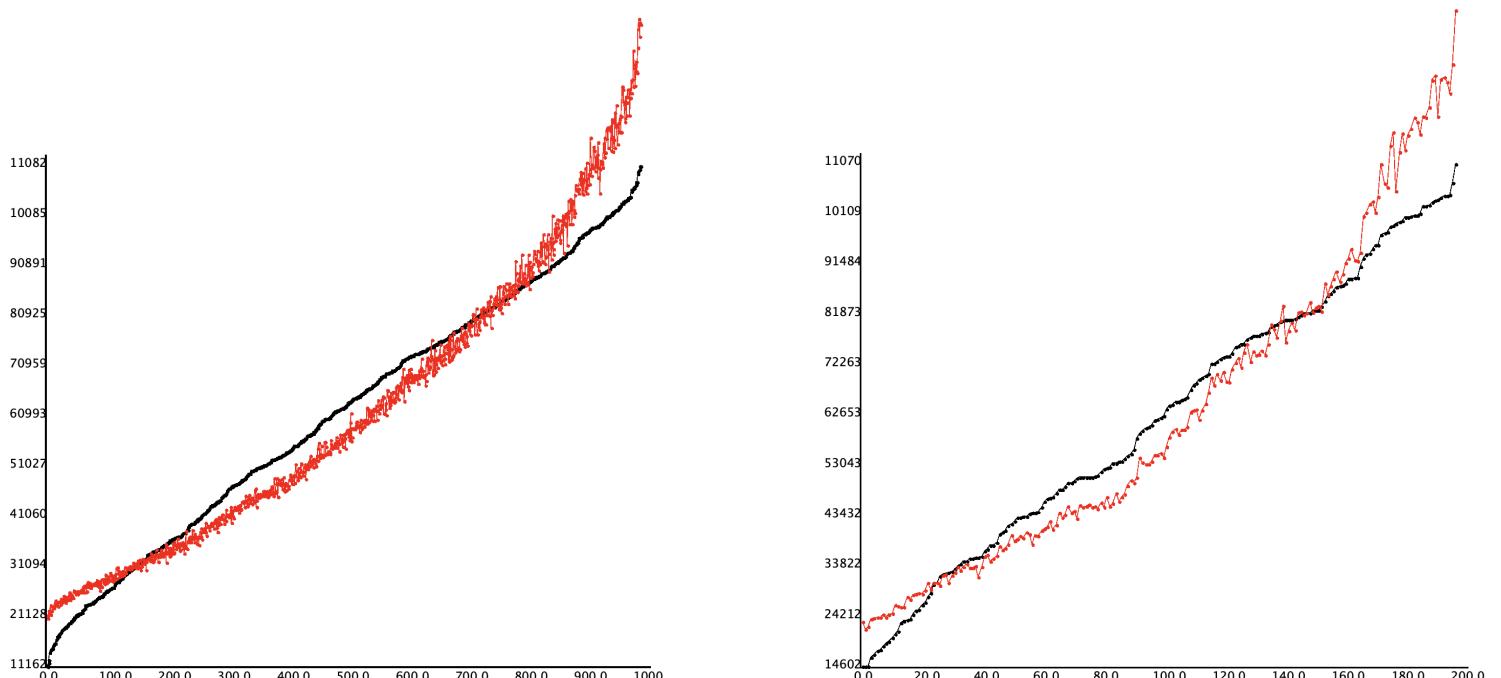


Figure 28: Scalation - House Price log1p

Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 $yy \text{ black/actual vs. } yp \text{ red/predicted}$

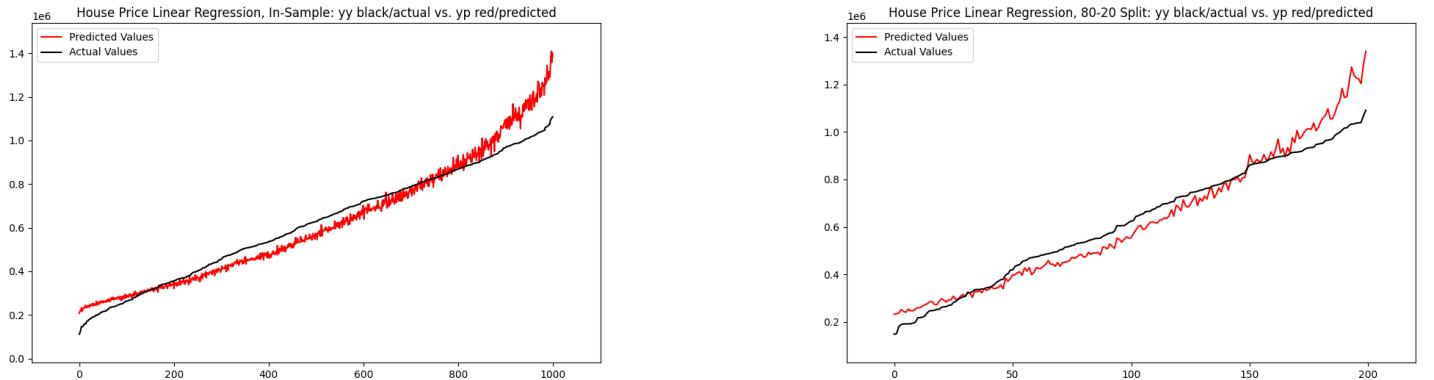


Figure 29: Statsmodels - House Price log1p

Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 $yy \text{ black/actual vs. } yp \text{ red/predicted}$

Table 17: Medical Cost (In-Sample): Sqrt, Log1p, and Box–Cox( $\lambda = 0.04$ )

Metric	sqrt	log1p	box-cox( $\lambda=0.04$ )
rSq	0.752657	0.522782	0.576265
rSqBar	0.751168	0.519909	0.573715
sst	1.96074e+11	1.96074e+11	1.96074e+11
sse	4.84975e+10	9.35702e+10	8.30835e+10
sde	6001.71	8358.44	7870.39
mse0	3.62463e+07	6.99329e+07	6.20953e+07
rmse	6020.49	8362.59	7880.06
mae	3613.90	4219.51	4052.90
smape	27.6903	26.2889	26.0851
m	1338.00	1338.00	1338.00
dfr	8.00000	8.00000	8.00000
df	1329.00	1329.00	1329.00
fStat	505.514	181.986	225.924
aic	-13525.1	-13964.7	-13885.2
bic	-13478.3	-13917.9	-13838.4

Table 18: Medical Cost (Validation): Sqrt, Log1p, and Box–Cox( $\lambda = 0.04$ )

Metric	sqrt	log1p	box-cox( $\lambda=0.04$ )
rSq	0.730154	0.571048	0.609124
rSqBar	0.728530	0.568466	0.606771
sst	4.06432e+10	4.06432e+10	4.06432e+10
sse	1.09674e+10	1.74340e+10	1.58865e+10
sde	6405.50	8088.13	7716.02
mse0	4.10764e+07	6.52957e+07	5.94998e+07
rmse	6409.08	8080.58	7713.61
mae	3839.79	4116.83	3998.83
smape	30.1151	27.3599	27.2984
m	267.000	267.000	267.000
dfr	8.00000	8.00000	8.00000
df	1329.00	1329.00	1329.00
fStat	449.505	221.156	258.882
aic	-2701.24	-2763.11	-2750.71
bic	-2668.95	-2730.83	-2718.42

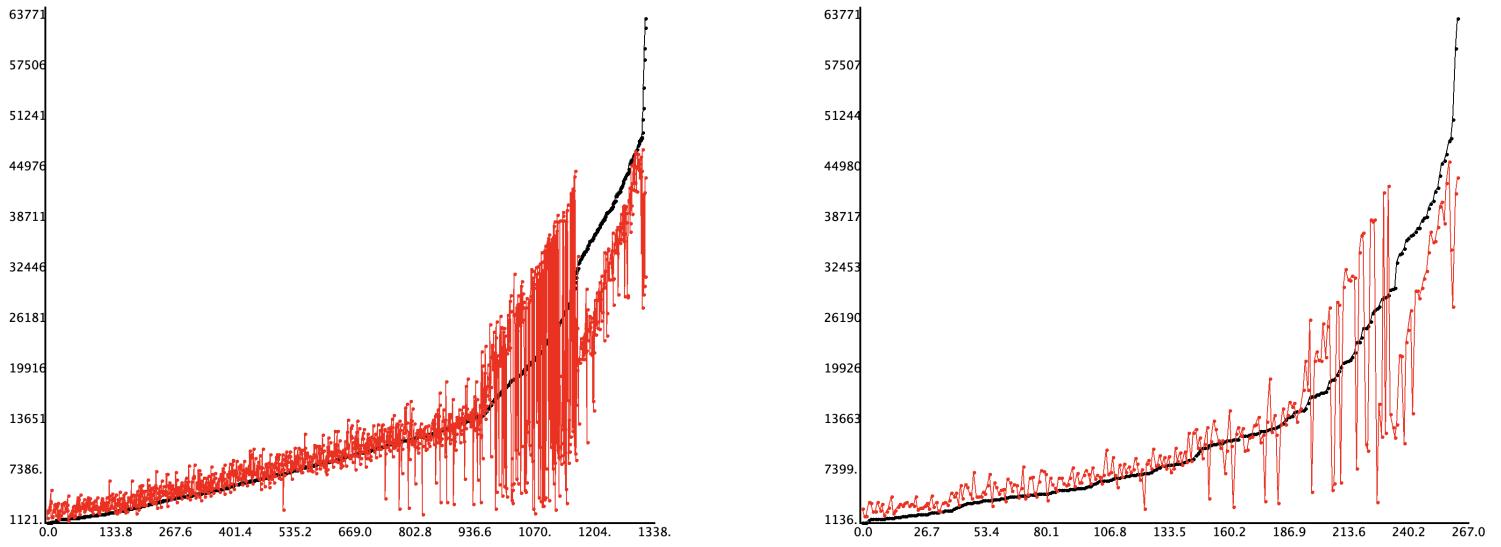


Figure 30: Scalation - Insurance Charges Sqrt

Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

Table 19: Auto MPG Regression with Square-Root Transformation

Table 20: Auto MPG Regression with Log1p Transformation

Table 21: Auto MPG Regression with Box-Cox Transformation

Table 22: Boston House Price Regression with Square-Root Transformation

Table 23: Boston House Price Regression with Log1p Transformation

Table 24: Boston House Price Regression with Box-Cox Transformation

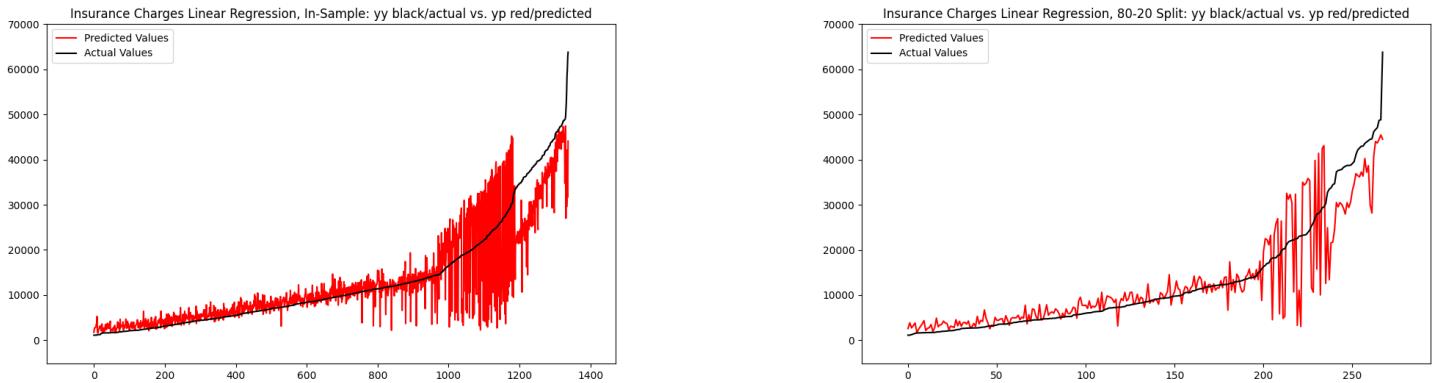


Figure 31: Statsmodels - Insurance Charges Sqrt

Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 $yy \text{ black/actual vs. } yp \text{ red/predicted}$

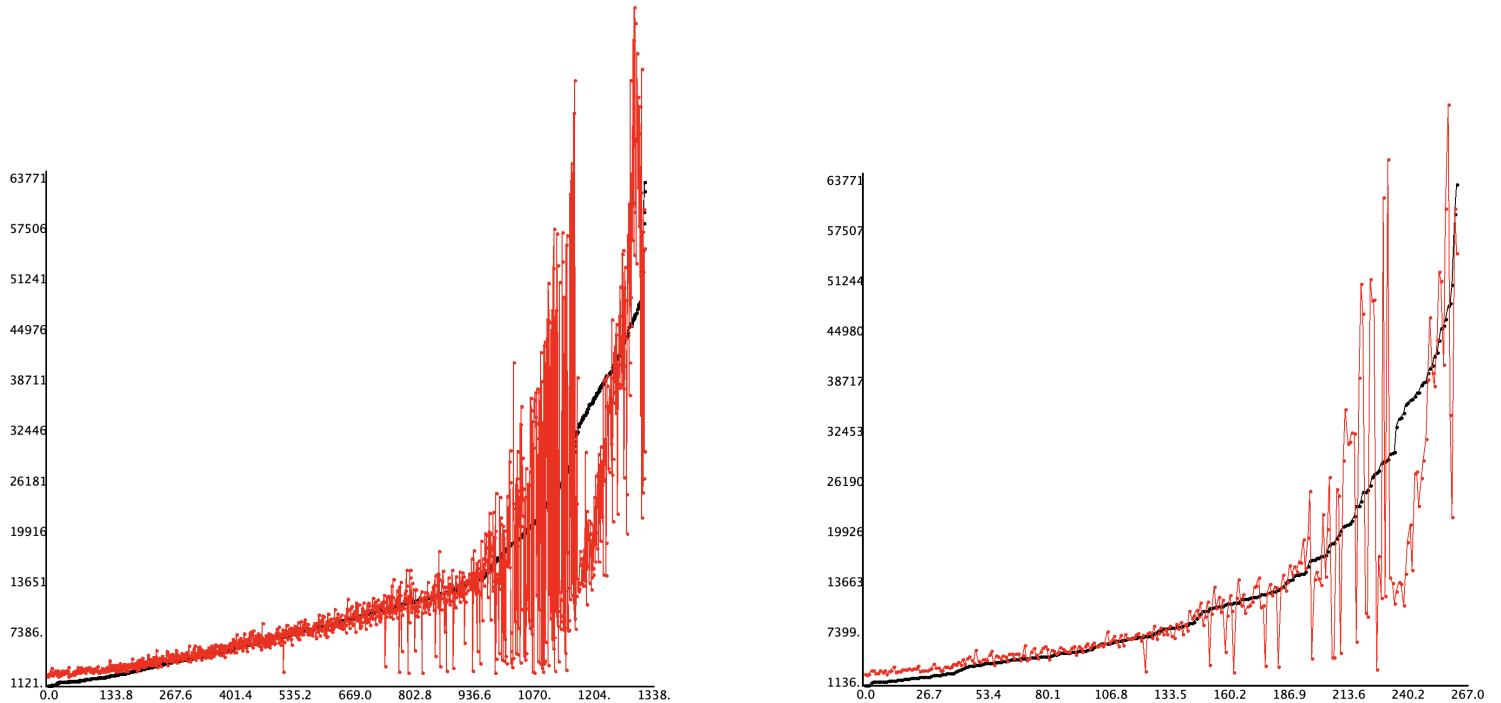


Figure 32: Scalation - Insurance Charges log1p

Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 $yy \text{ black/actual vs. } yp \text{ red/predicted}$

Table 25: Medical Cost Regression with Square-Root Transformation

Table 26: Medical Cost Regression with Log1p Transformation

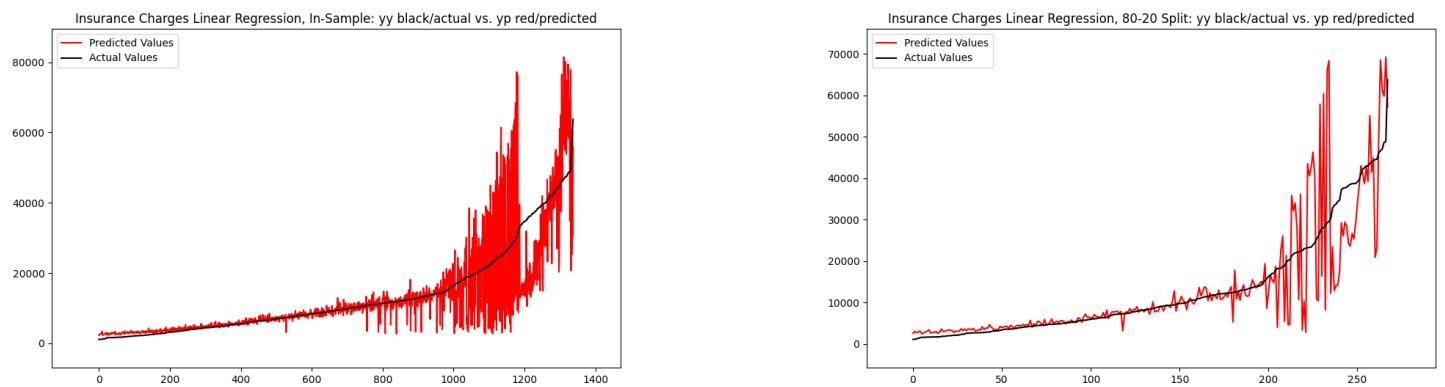


Figure 33: Statsmodels - Insurance Charges log1p  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 $yy \text{ black/actual vs. } yp \text{ red/predicted}$