

# Project 1

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

February 2026

## 1 Introduction

In this project, we will explore three different data sets using multiple linear regression, ridge regression, lasso, and transformed regression (specifically sqrt, log1p, and Box-Cox). The three data sets we will be using are:

- **Auto MPG:** This data set contains information about different cars. The goal is predict the miles per gallon (MPG) given the other characteristics of the car. The data set can be found at <https://archive.ics.uci.edu/dataset>.
- **House Prices:** This data set contains information about different houses. The goal is to predict the price of the house given the other characteristics. The data set can be found at <https://www.kaggle.com/datasets/prokshitha/home-value-insights>.
- **Insurance Charges:** This data set contains information about different people. The goal is to predict a persons insurance charges given their other characteristics. The data set can be found at <https://www.kaggle.com/datasets/mirichoi0218/insurance>.

The report will be organized as follows, Section 2 will be focused on linear regression, Section 3 ridge regression, Section 4 lasso, Section 5 transformed regression, and Section 6 will be comparing all of the models.

In Section 2, in-sample, out-of-sample, 5-fold cross validation, and feature selection results are given for multiple linear regression on all three data sets. The results come from both scalation in scala and statsmodels in python.

## 2 Linear Regression

In this section, multiple linear regression is applied to the Auto MPG, House Prices, and Insurance Charges data sets. We evaluated the model using in-sample, out-of-sample, and cross validation. Finally, we applied forward selection, backward elimination, and stepwise selection to identify which features are optimal for explanation of the response variable.

### 2.1 Auto MPG

Table 1 presents the quality of fit metrics for the in-sample and out-of-sample evaluations using scalation, while Table 2 presents the same metrics using statsmodels. We can see that regression is performing decently, and that the main statistics are similar for both scalation and mathstats.

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

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

Figure 1 shows that plots for the predicted  $y$ -values vs. the actual  $y$ -values from scalation, while Figure 2 show the same from mathstats. Again the results are very similar.

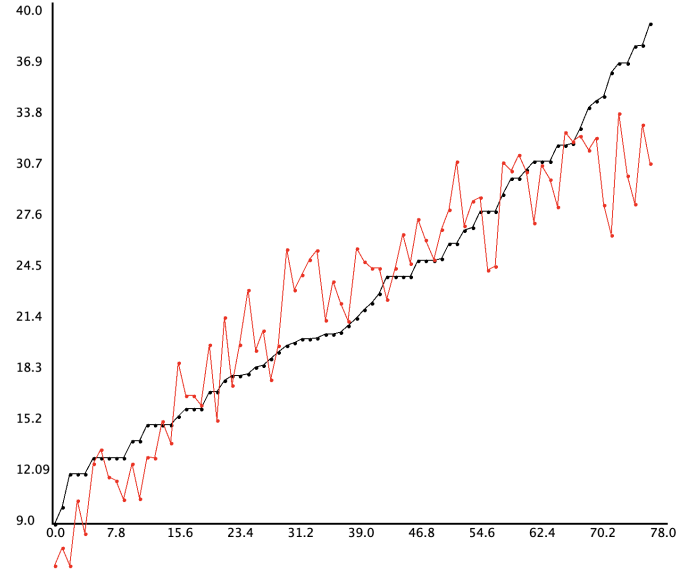
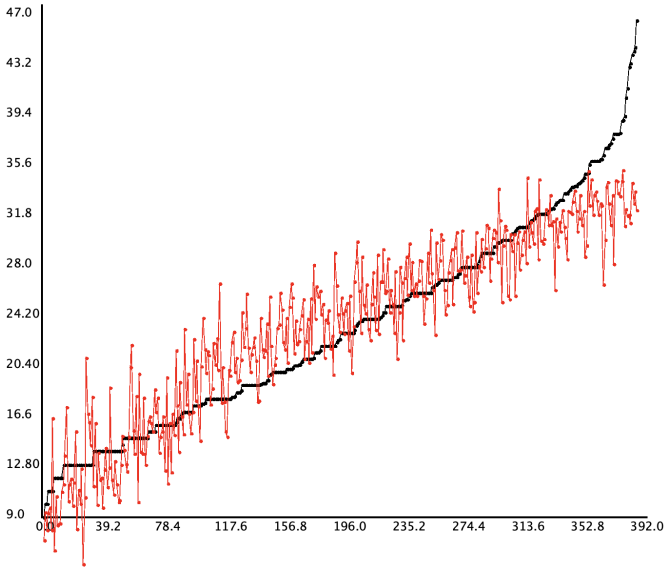


Figure 1: Scalation - Auto MPG Regression  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

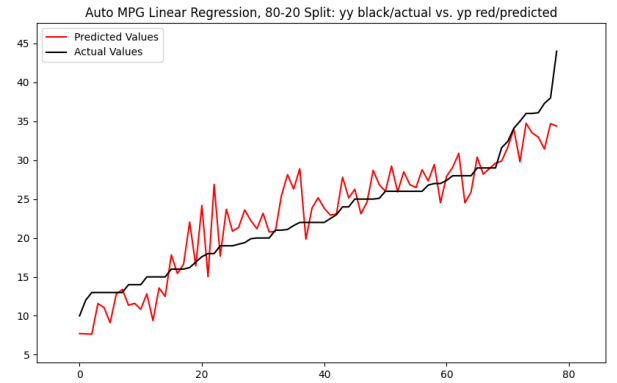
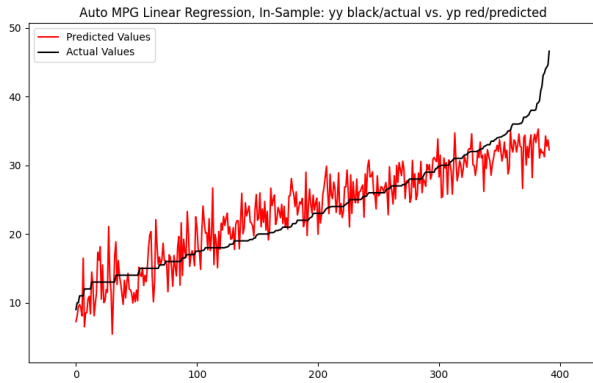


Figure 2: Statsmodels - Auto MPG Regression  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

Next, we performed 5 fold cross validation. Tables 3 and 4 show the resulting quality of fit metrics from scalation and statsmodels respectively. Again we see that the results are similar.

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

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

Finally, we apply forward selection, backward elimination, and stepwise selection to determine which variables best explain response variable. Figure 3 shows plots for  $R^2$ ,  $\overline{R}^2$ , sMAPE,  $R^2$  cv, and AIC vs.  $n$  (the number of variables selected) when utilizing forward selection. From the left plot, it is clear that only 2 variables are needed to explain the response variable (even just 1 variable is not too bad either). Those variables 2 variables are weight and modelyear. The order in which the variables were chosen was 1. weight, 2. modelyear, 3. cylinders, 4. acceleration, 5. horsepower, and 6. displacment.

One interesting thing to note is how the graph of AIC vs.  $n$  is always increasing.

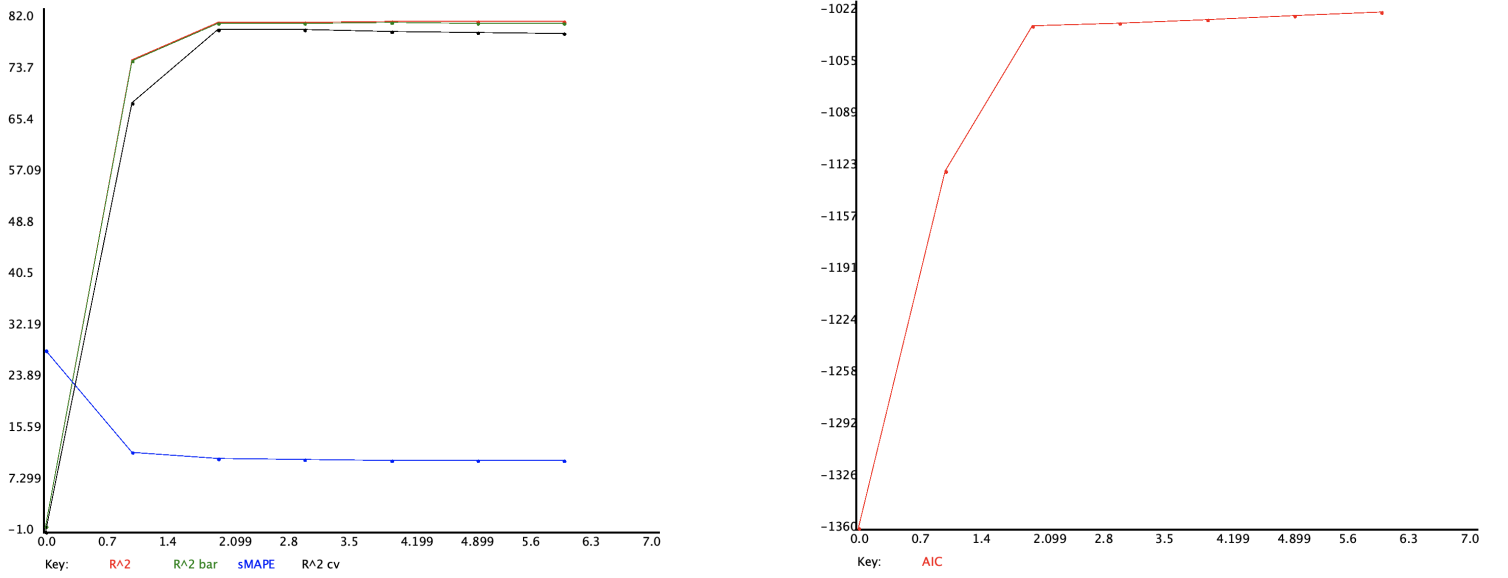


Figure 3: Scalation - Auto MPG Regression Forward Selection

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

Right: aic vs.  $n$

Figure 4 shows plots for  $R^2$ ,  $\overline{R^2}$ , sMAPE,  $R^2$  cv, and AIC vs.  $n$  (the number of variables selected) when utilizing backward elimination. From the left plot, it is again clear that only 2 variables are needed to explain the response variable (and again even just 1 variable is not too bad either). Here, both backward and forward selection agree with each other in their selection of variables, so if you were to remove one variable at a time, in order you would remove 1. displacement, 2. horsepower, 3. acceleration, 4. cylinders, 5. modelyear, and 6. weight. This is the reverse order of forward selection, meaning that they agree.

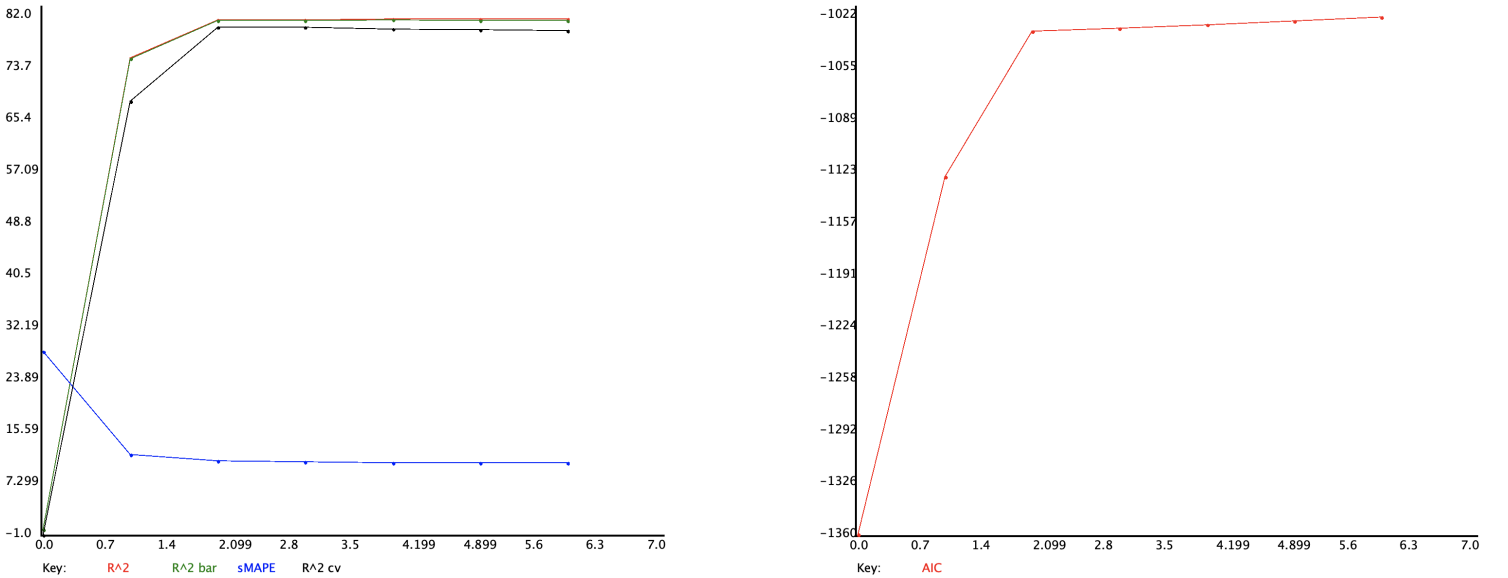


Figure 4: Scalation - Auto MPG Regression Backward Elimination

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

Right: aic vs.  $n$

Last, Figure 5 shows plots for  $R^2$ ,  $\overline{R^2}$ , sMAPE,  $R^2$  cv, and AIC vs.  $n$  (the number of variables selected) when utilizing stepwise selection. Here, since stepwise can move forward and backwards which results in multiple different metric evaluations for models that use the same number of variables, (I believe) the evaluation that is plotted for  $n$  variables is the best evaluation that stepwise selection found for all the models with  $n$  variables that it tested. Despite this, we still find that if we want to get

the best models recommended from stepwise, we should add in the following order, 1. weight, 2. modelyear, 3. cylinders, 4. acceleration, and 5. horsepower. Note that this almost agrees with forward selection and backward elimination, but we do not add displacement as stepwise decided that moving was not worthwhile.

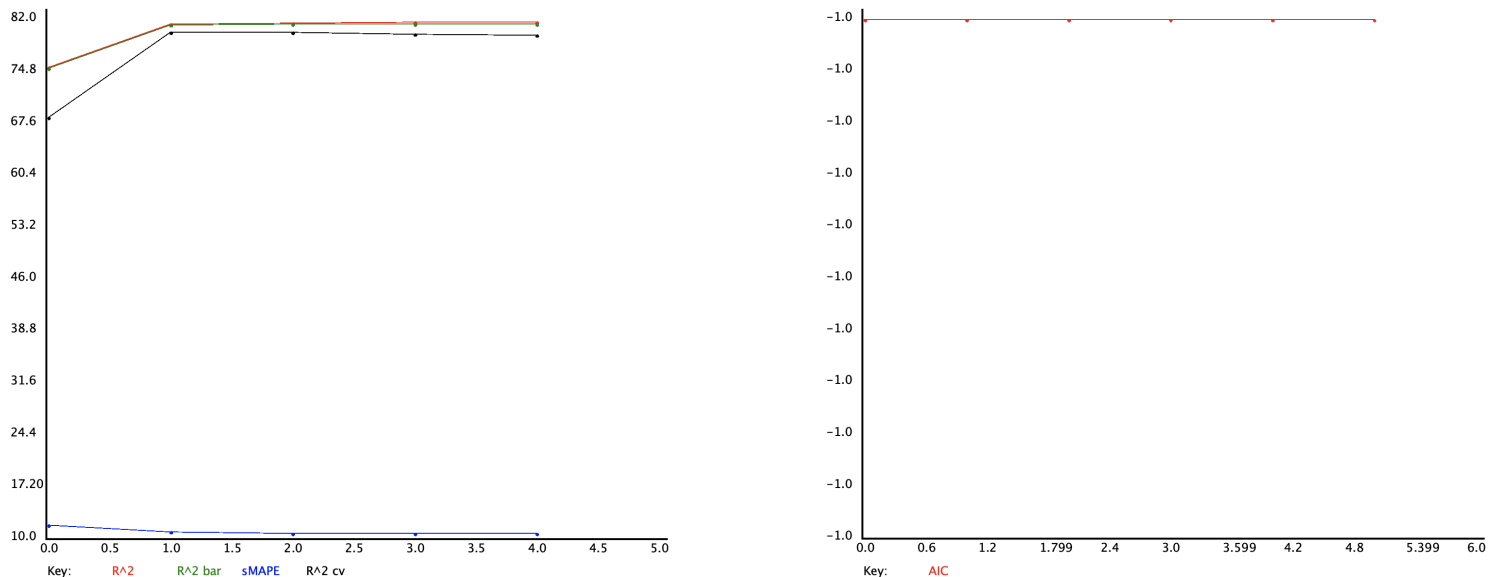


Figure 5: Scalation - Auto MPG Regression Stepwise Selection

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

Right: aic vs.  $n$

## 2.2 House Prices

Table 5 presents the quality of fit metrics for the in-sample and out-of-sample evaluations using scalation, while Table 6 presents the same metrics using statsmodels. We can see that regression is performing extremely well, almost perfectly capturing the data. The main statistics are similar for both scalation and mathstats.

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 shows that plots for the predicted  $y$ -values vs. the actual  $y$ -values from scalation, while Figure 7 show the same from mathstats. Again the results are very similar.

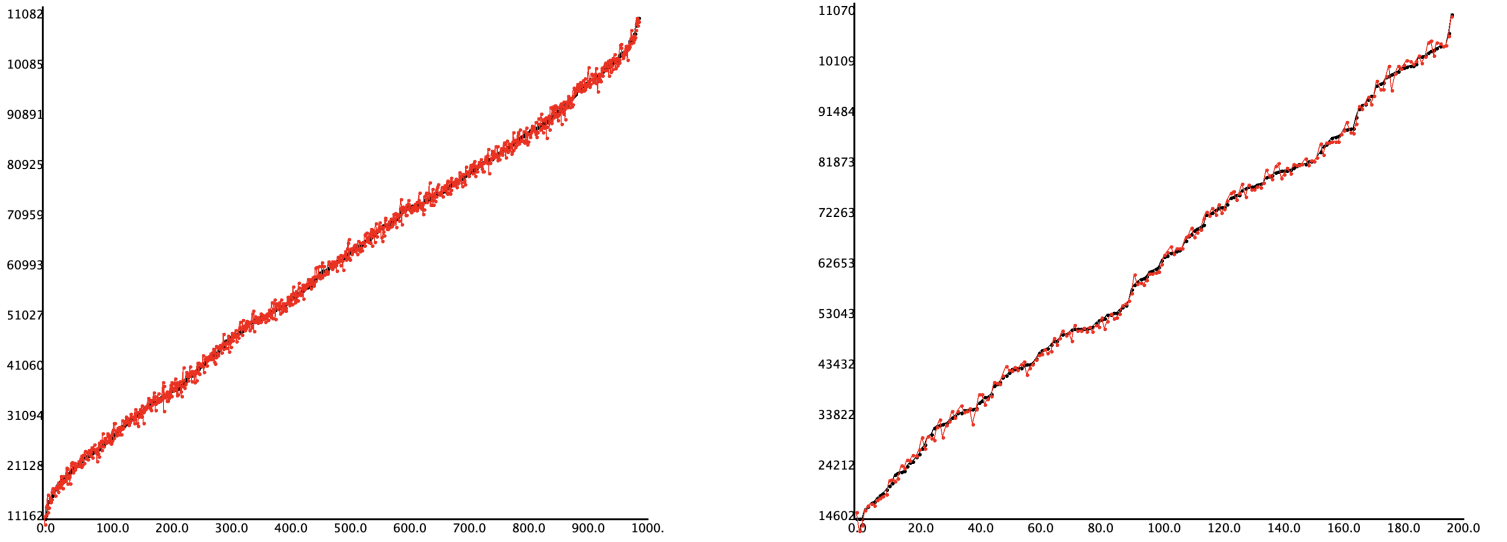


Figure 6: Scalation - House Price Regression  
Left: In Sample Predictions  
Right: 80-20 Out of Sample Predictions  
yy black/actual vs. yp red/predicted

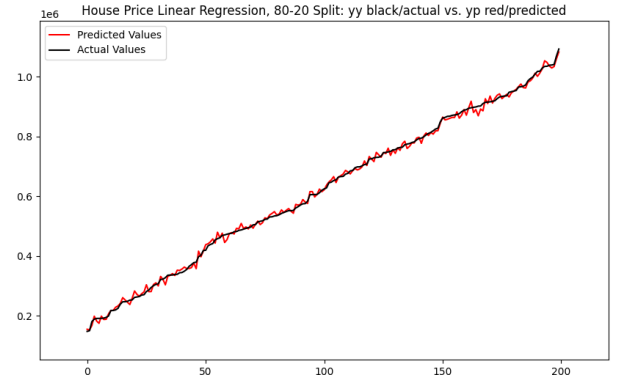
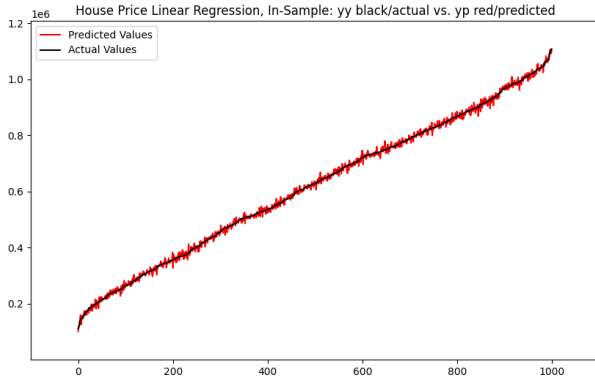


Figure 7: Statsmodels - House Price Regression  
Left: In Sample Predictions  
Right: 80-20 Out of Sample Predictions  
yy black/actual vs. yp red/predicted

Next, we performed 5 fold cross validation. Tables 7 and 8 show the resulting quality of fit metrics from scalation and statsmodels respectively. Again we see that the results are similar.

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

Finally, we apply forward selection, backward elimination, and stepwise selection to determine which variables best explain response variable. Figure 8 shows plots for  $R^2$ ,  $\overline{R}^2$ , sMAPE,  $R^2$  cv, and AIC vs.  $n$  (the number of variables selected) when utilizing forward selection. From the left plot, it is clear that only 1 variable is needed to explain the response variable. The order in which the variables were chosen was 1. Square Footage, 2. Year Built, 3. Lot Size, 4. Num Bedrooms, 5. Num Bathrooms, and 6. Garage Size.

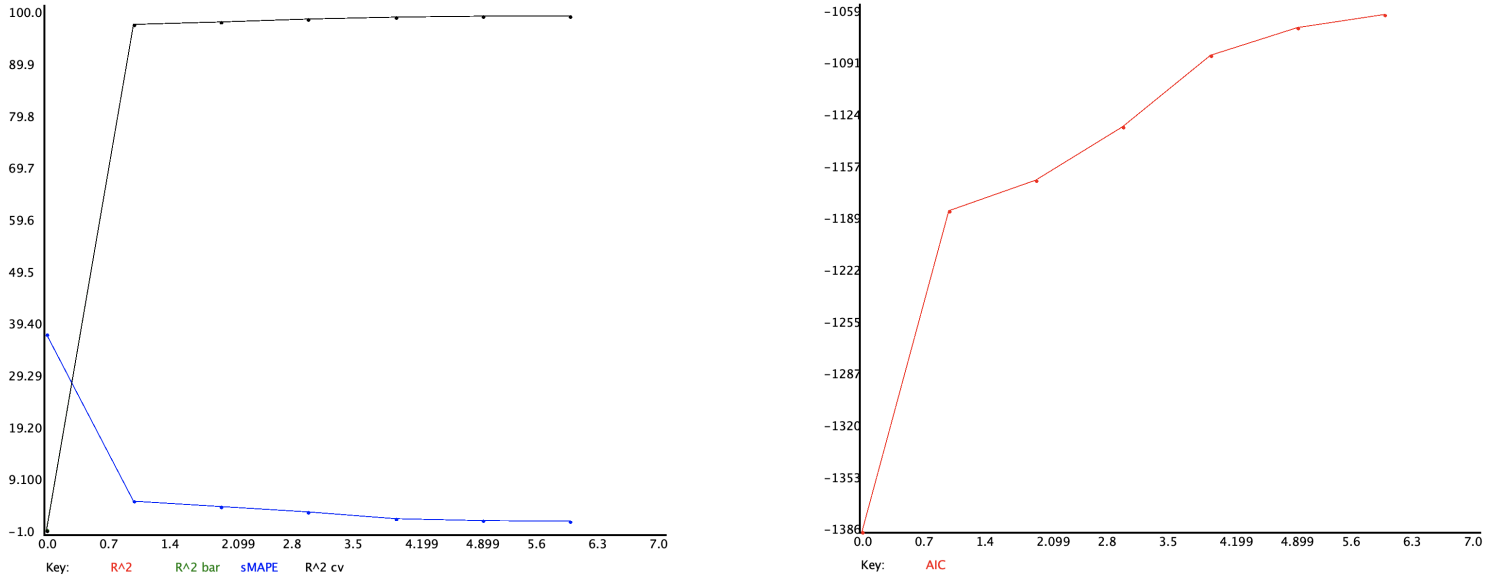


Figure 8: Scalation - House Prices Regression Forward Selection

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

Figure 9 shows plots for  $R^2$ ,  $\overline{R}^2$ , sMAPE,  $R^2$  cv, and AIC vs.  $n$  (the number of variables selected) when utilizing backward elimination. From the left plot, it is again clear that only 1 variable is needed to explain the response variable. Here, both backward and forward selection agree with each other in their selection of variables, so if you were to remove one variable at a time, in order you would remove 1. Garage Size, 2. Num Bathrooms, 3. Num Bedrooms, 4. Lot Size, 5. Year Built, and 6. Square Footage. This is the reverse order of forward selection, meaning that they agree.

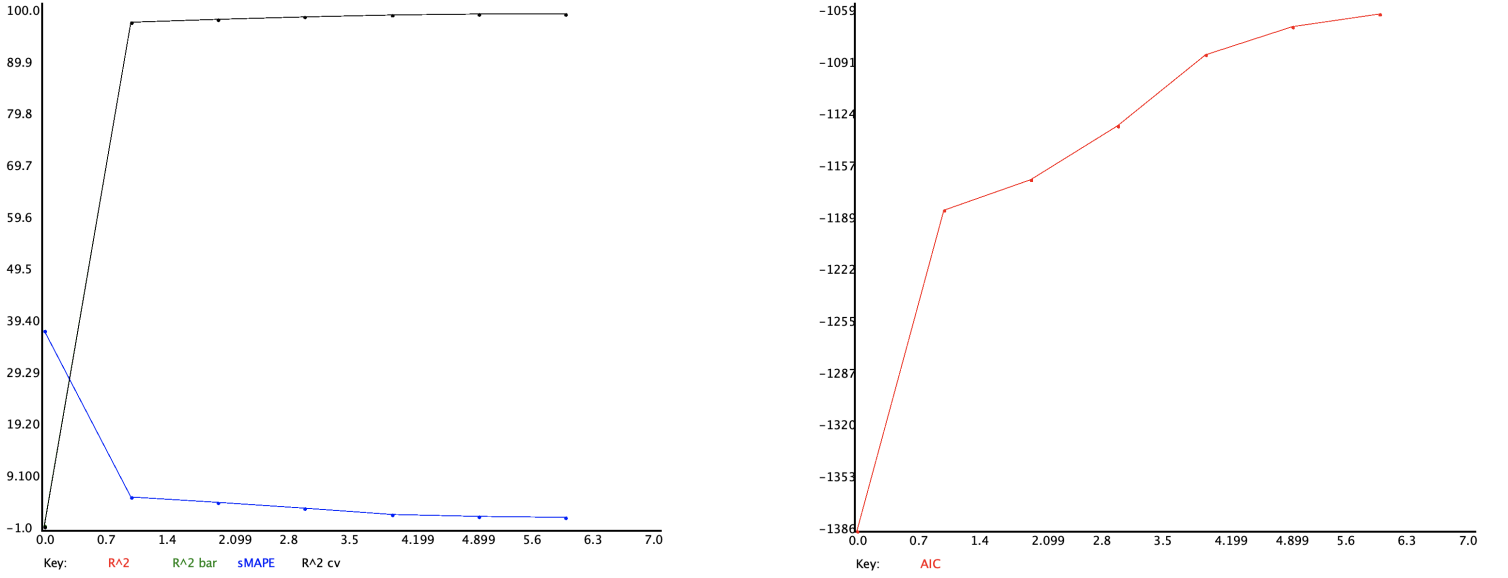


Figure 9: Scalation - House Prices Regression Backward Elimination

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

Right: aic vs.  $n$

Last, Figure 10 shows plots for  $R^2$ ,  $\overline{R^2}$ , sMAPE,  $R^2$  cv, and AIC vs.  $n$  (the number of variables selected) when utilizing stepwise selection. We still find that if we want to get the best models recommended from stepwise, we should add in the following order, 1. Square Footage, 2. Year Built, 3. Lot Size, 4. Num Bedrooms, and 5. Num Bathrooms. Note that this almost agrees with forward selection and backward elimination, but we do not add Garage Size as stepwise decided that moving was not worthwhile.

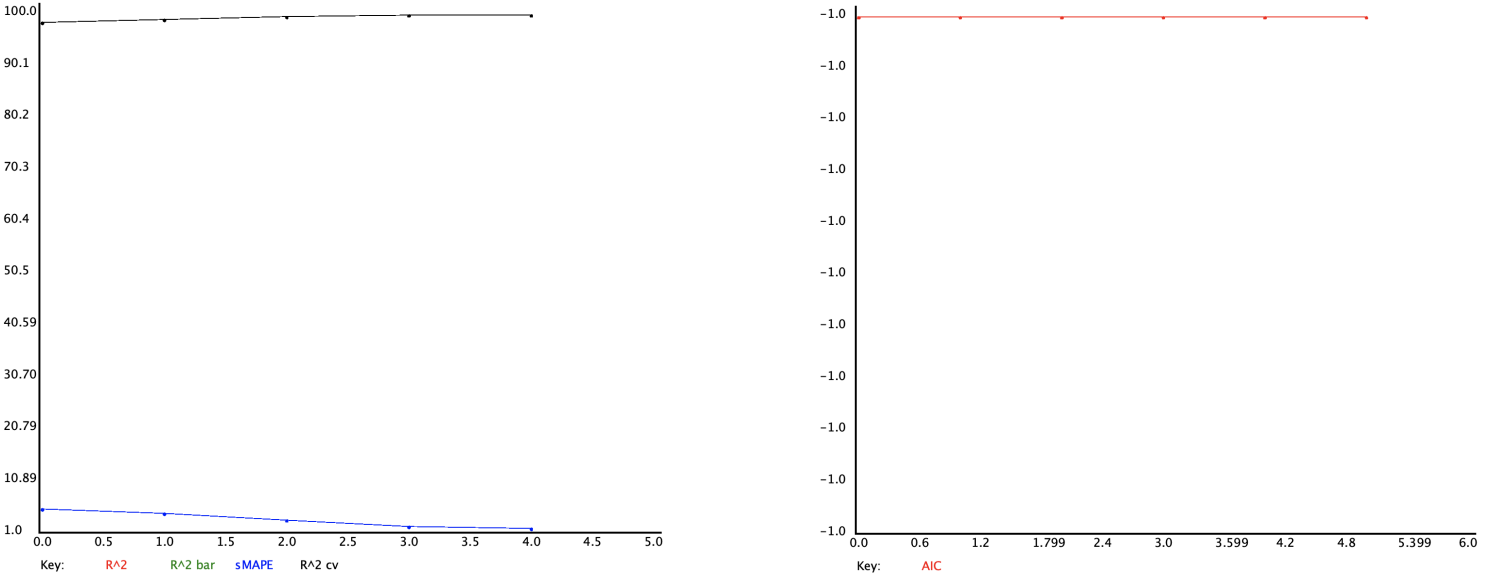


Figure 10: Scalation - House Prices Regression Stepwise Selection

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

Right: aic vs.  $n$

## 2.3 Insurance Charges

Table 9 presents the quality of fit metrics for the in-sample and out-of-sample evaluations using scalation, while Table 10 presents the same metrics using statsmodels. We can see that regression is performing okay, and that the main statistics are similar for both scalation and mathstats.

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 shows that plots for the predicted  $y$ -values vs. the actual  $y$ -values from scalation, while Figure 12 show the same from mathstats. Again the results are very similar.

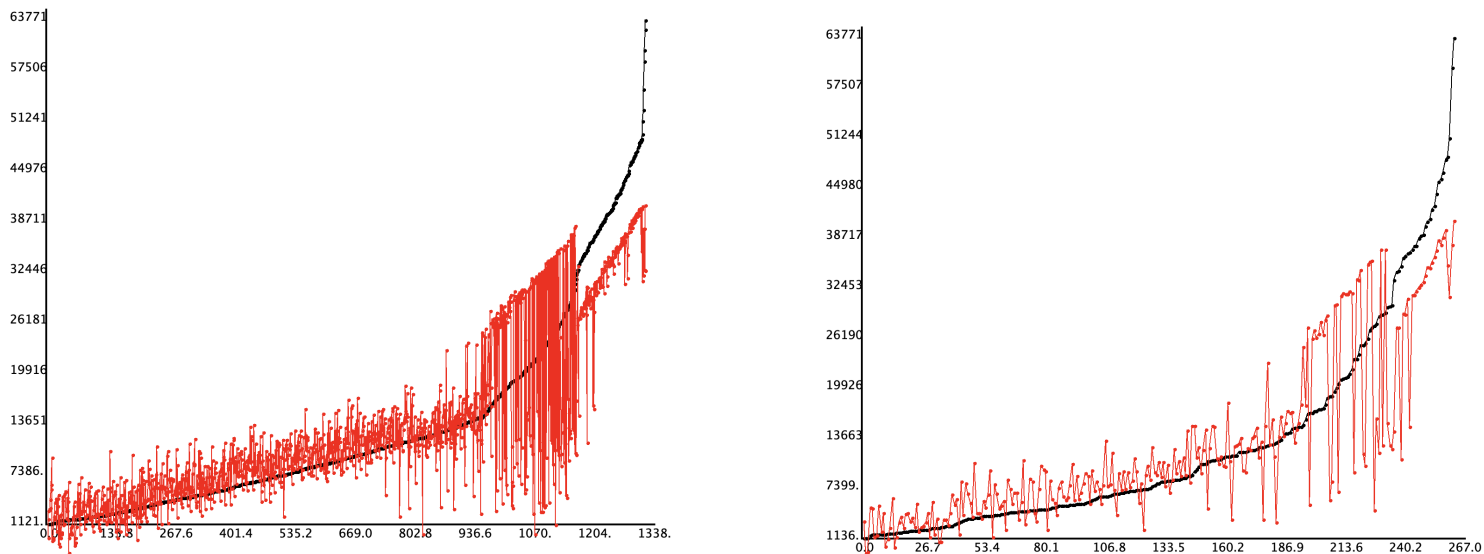


Figure 11: Scalation - Insurance Charges Regression

Left: In Sample Predictions

Right: 80-20 Out of Sample Predictions

yy black/actual vs. yp red/predicted

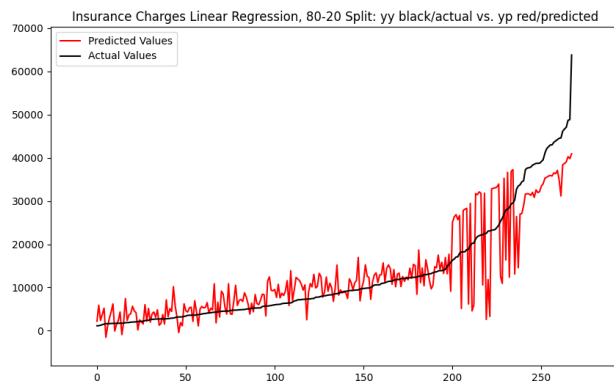
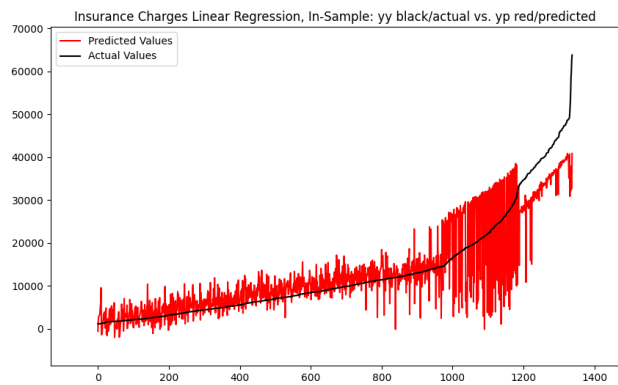


Figure 12: Statsmodels - Insurance Charges Regression

Left: In Sample Predictions

Right: 80-20 Out of Sample Predictions

yy black/actual vs. yp red/predicted

Next, we performed 5 fold cross validation. Tables 11 and 12 show the resulting quality of fit metrics from scalation and statsmodels respectively. Again we see that the results are similar.

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

Finally, we apply forward selection, backward elimination, and stepwise selection to determine which variables best explain response variable. Figure 13 shows plots for  $R^2$ ,  $\bar{R}^2$ , sMAPE,  $R^2$  cv, and AIC vs.  $n$  (the number of variables selected) when utilizing forward selection. From the left plot, it is clear that only  $s$  variables are needed to explain the response variable. The order in which the variables were chosen was 1. smoker\_yes, 2. age, 3. bmi, 4. children, 5. region\_southeast, 6. sex\_male, and 7. region\_northwest

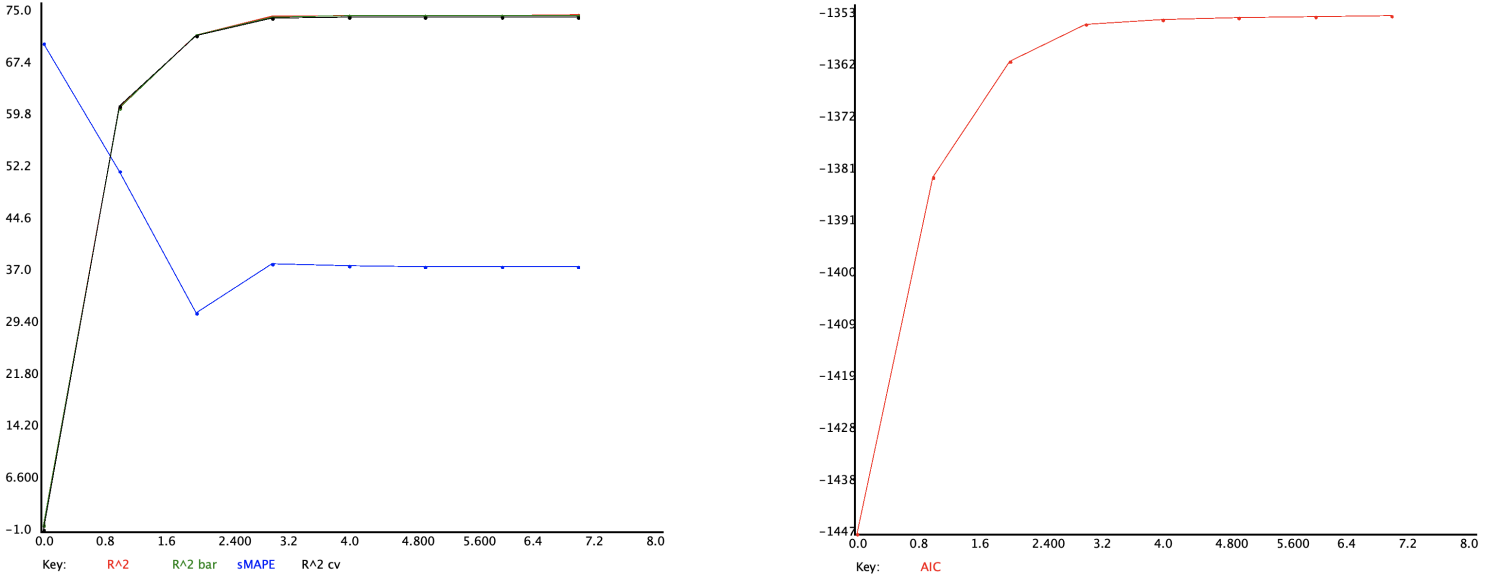


Figure 13: Scalation - Insurance Charges Regression Forward Selection

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

Right: aic vs.  $n$

Figure 14 shows plots for  $R^2$ ,  $\overline{R^2}$ , sMAPE,  $R^2$  cv, and AIC vs.  $n$  (the number of variables selected) when utilizing backward elimination. From the left plot, it is again clear that only 2 variables are needed to explain the response variable. Here, both backward and forward selection agree with each other in their selection of variables, so if you were to remove one variable at a time, in order you would remove 1. region\_northwest, 2. sex\_male, 3. region\_southeast, 4. children, 5. bmi, 6. age, and 7. smoker\_yes. This is the reverse order of forward selection, meaning that they agree.

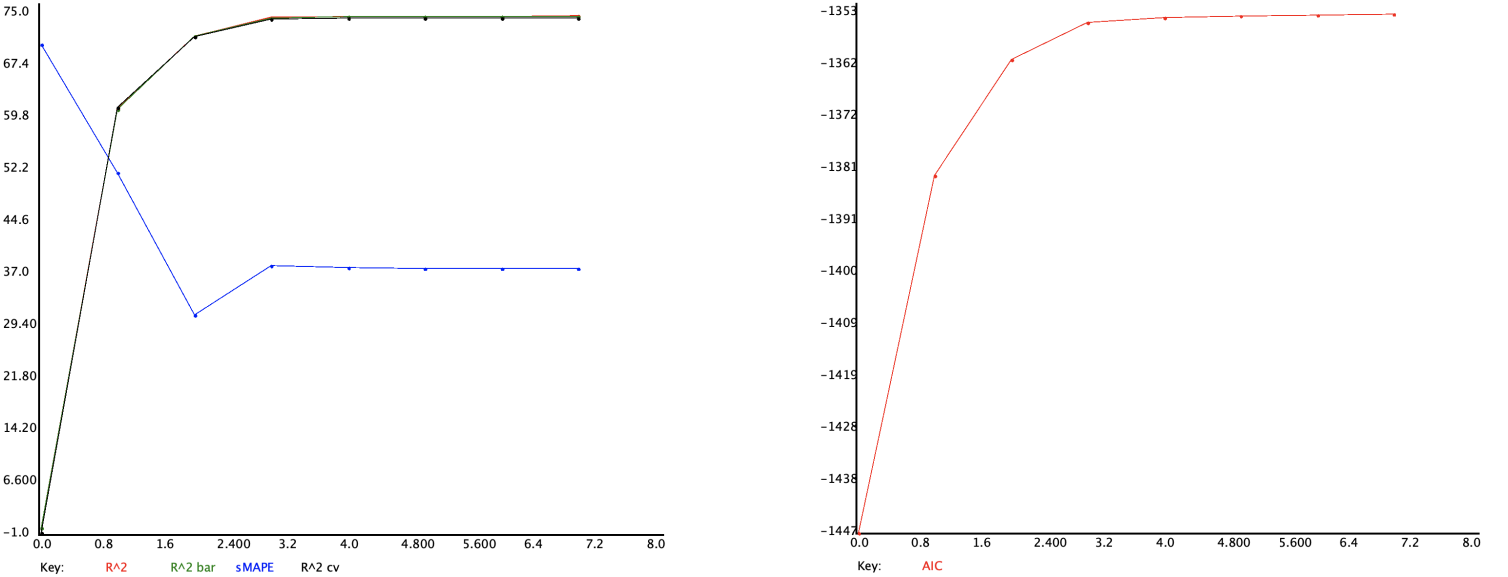


Figure 14: Scalation - Insurance Charges Regression Backward Elimination

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

Right: aic vs.  $n$

Last, Figure 15 shows plots for  $R^2$ ,  $\overline{R^2}$ , sMAPE,  $R^2$  cv, and AIC vs.  $n$  (the number of variables selected) when utilizing stepwise selection. We still find that if we want to get the best models recommended from stepwise, we should add in the following order, 1. smoker\_yes, 2. age, 3. bmi, 4. children, 5. region\_southeast, and 6. sex\_male. Note that this almost agrees with forward selection and backward elimination, but we do not add region\_northwest as stepwise decided that moving was not worthwhile.

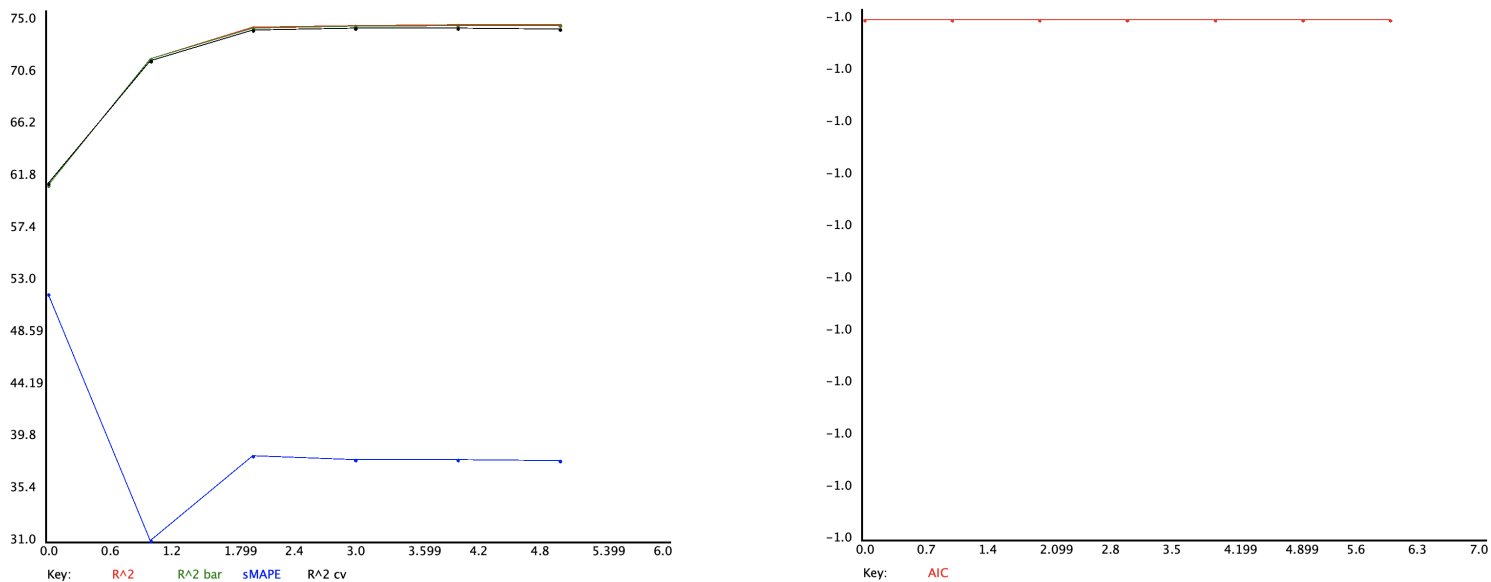


Figure 15: Scalation - Insurance Charges Regression Stepwise Selection

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

Right: aic vs.  $n$

### 3 Ridge Regression

#### 3.1 Auto MPG

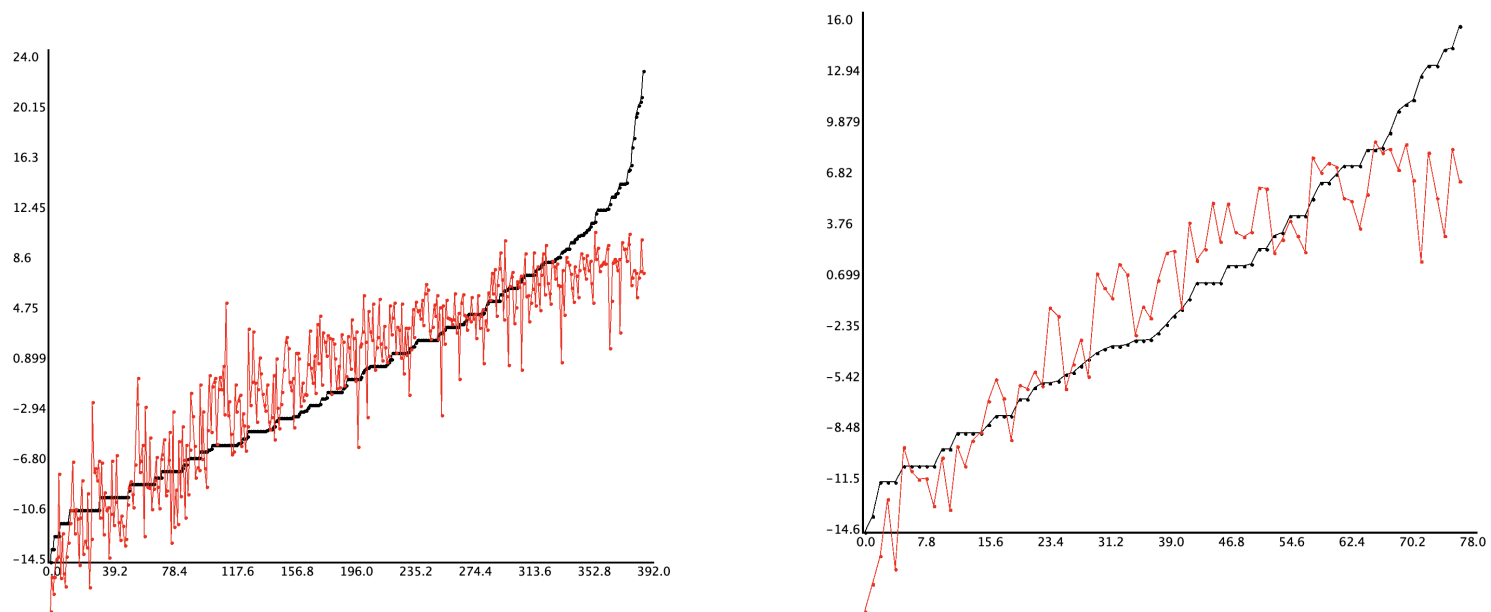


Figure 16: Scalation - Auto MPG Ridge

Left: In Sample Predictions

Right: 80-20 Out of Sample Predictions

$y$  black/actual vs.  $yp$  red/predicted

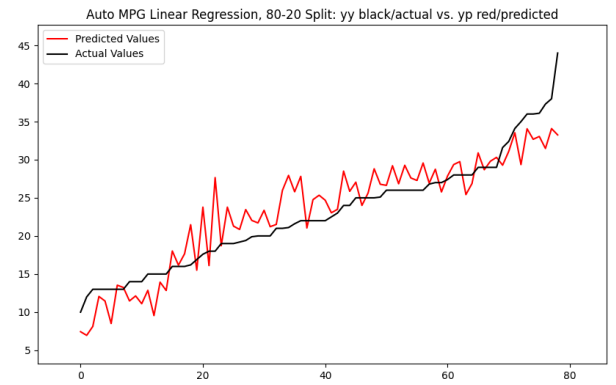
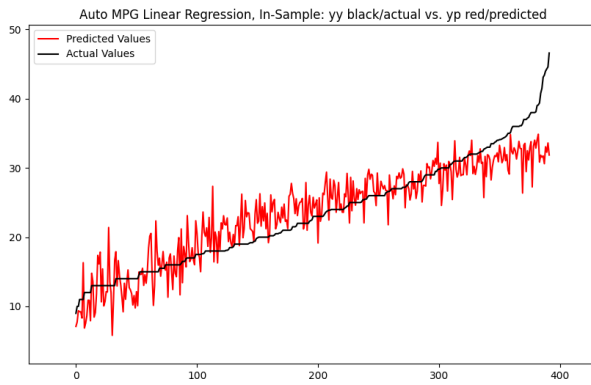


Figure 17: Statsmodels - Auto MPG Ridge  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

## 3.2 House Prices

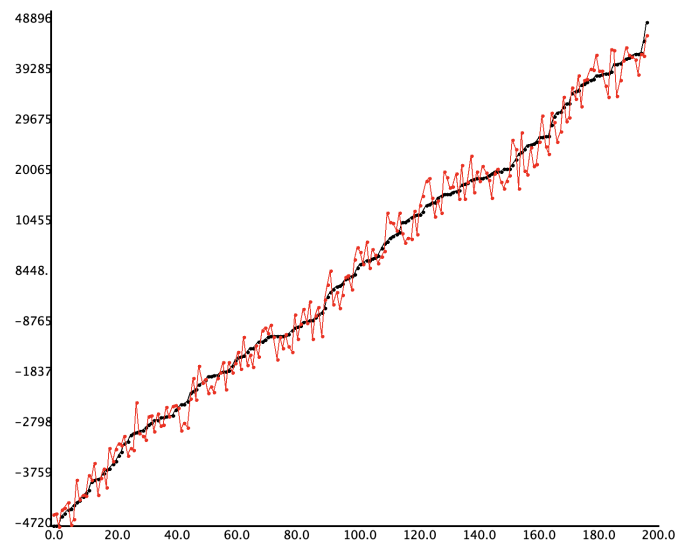
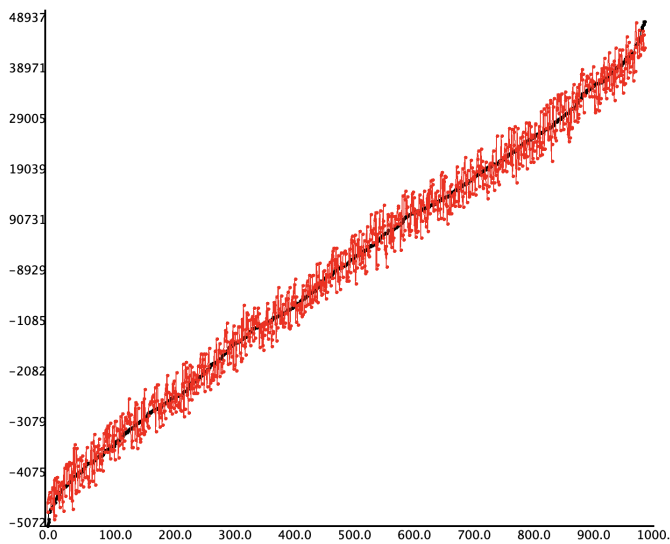


Figure 18: Scalation - House Price Ridge  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted



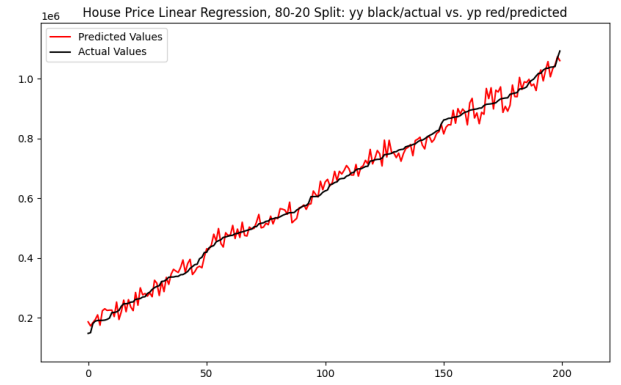
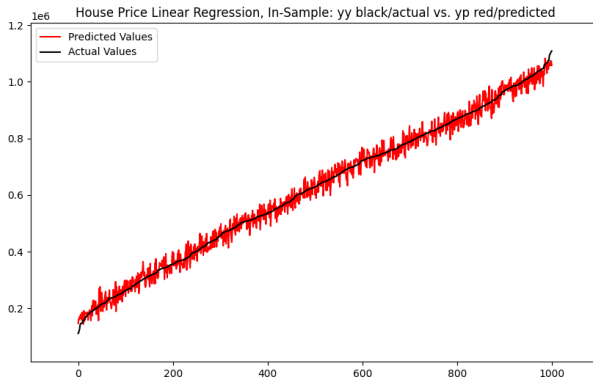


Figure 19: Statsmodels - House Price Ridge  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

### 3.3 Insurance Charges

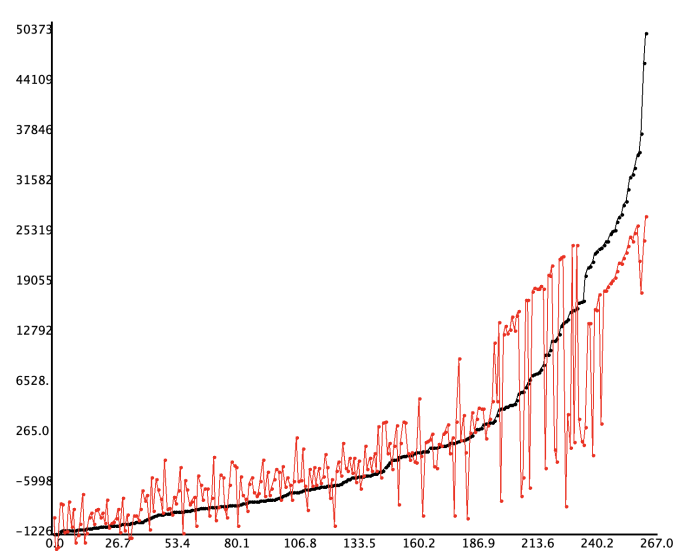
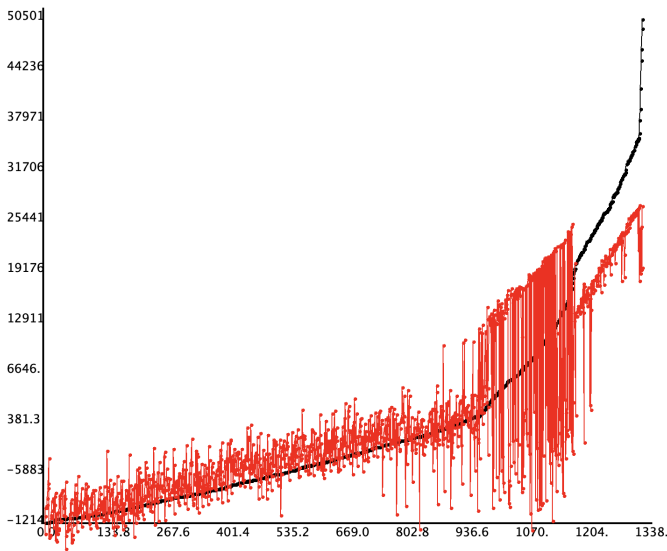


Figure 20: Scalation - Insurance Charges Ridge  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

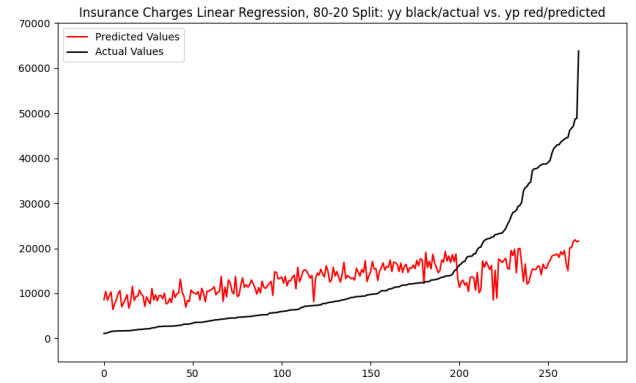
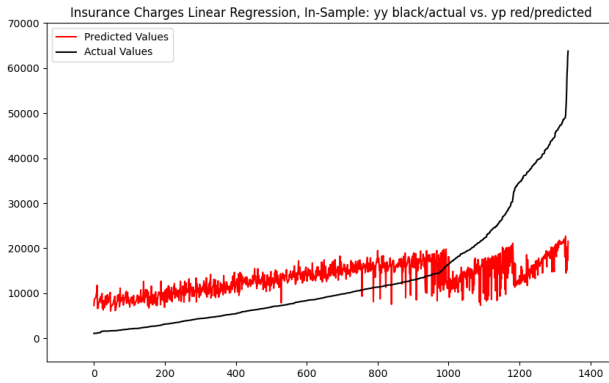


Figure 21: Statsmodels - Insurance Charges Ridge  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

## 4 Lasso

### 4.1 Auto MPG

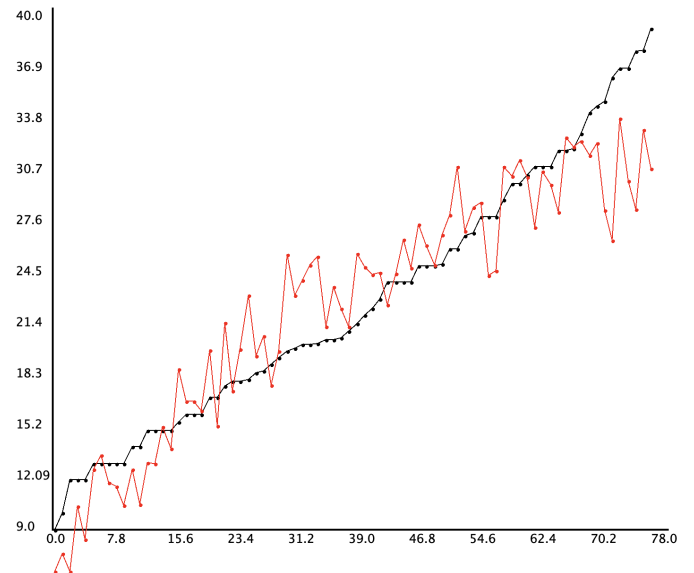
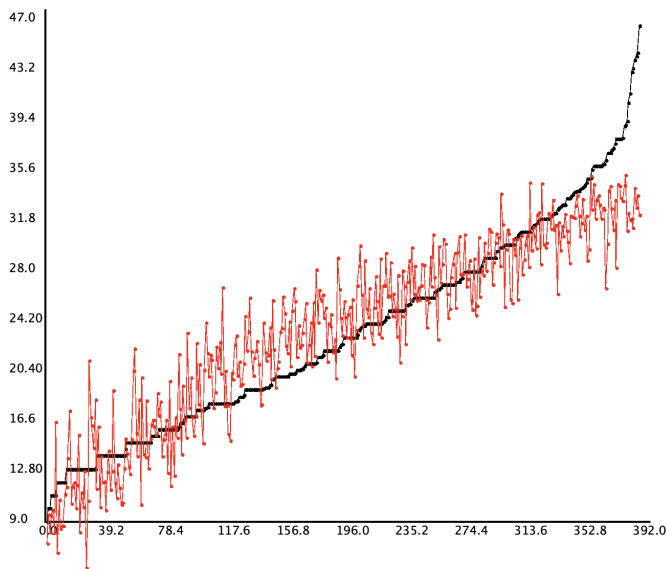


Figure 22: Scalation - Auto MPG Lasso  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

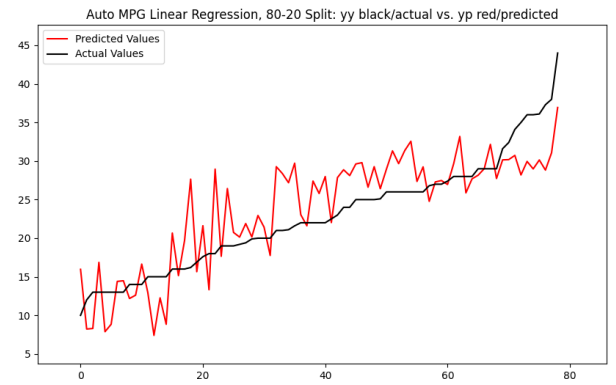
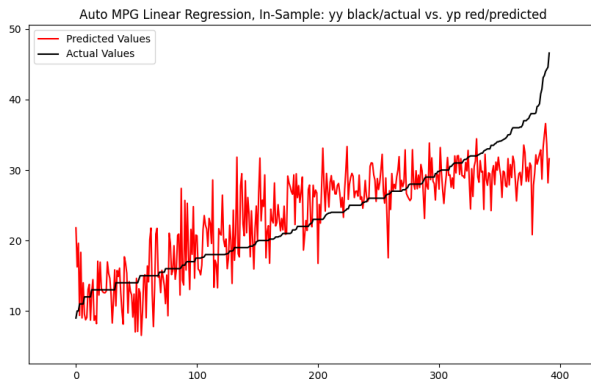


Figure 23: Statsmodels - Auto MPG Lasso  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

## 4.2 House Prices

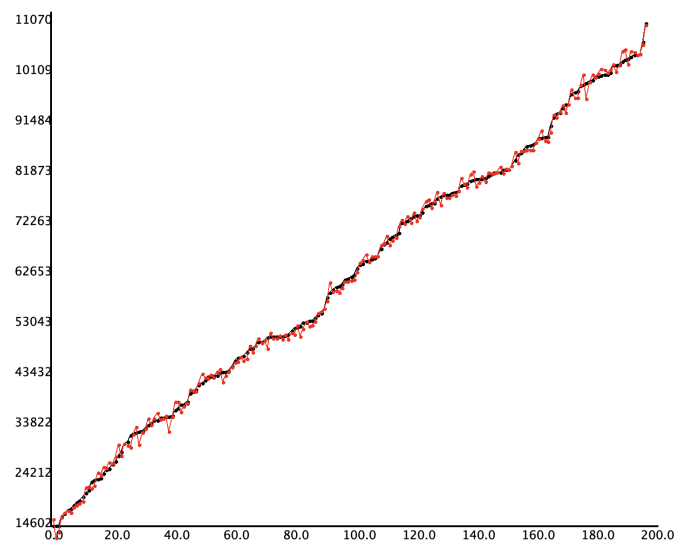
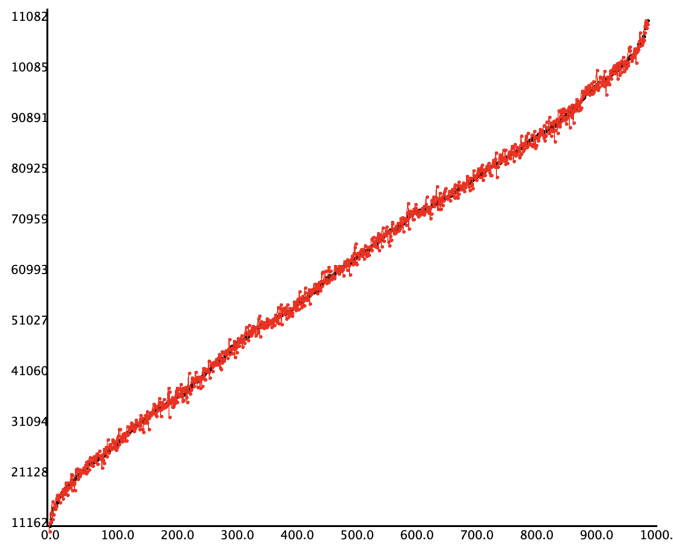


Figure 24: Scalation - House Price Lasso  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

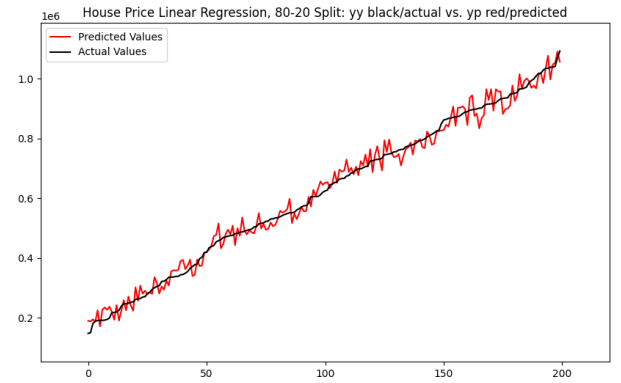
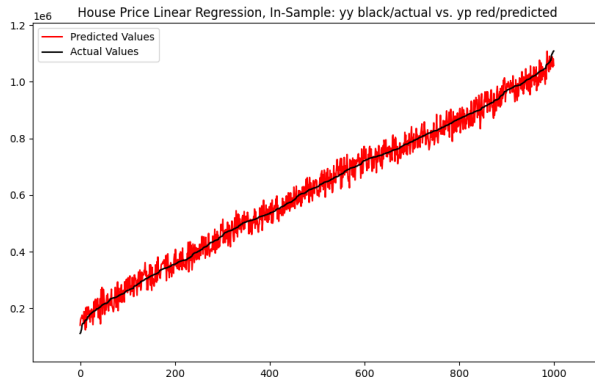


Figure 25: Statsmodels - House Price Lasso  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

### 4.3 Insurance Charges

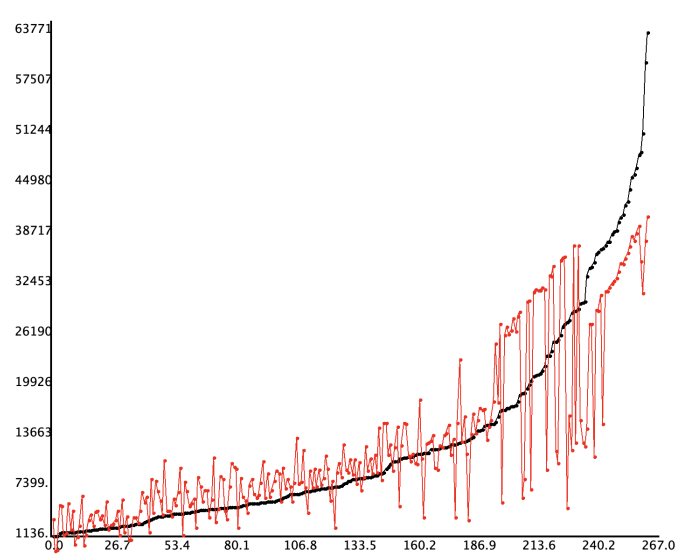
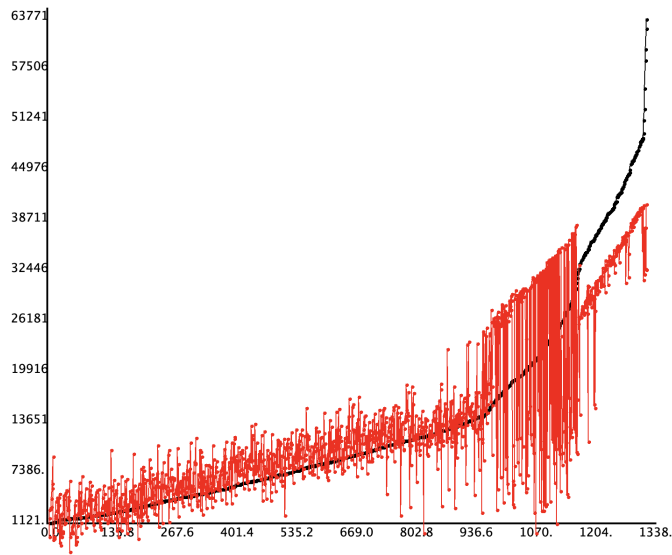


Figure 26: Scalation - Insurance Charges Lasso  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

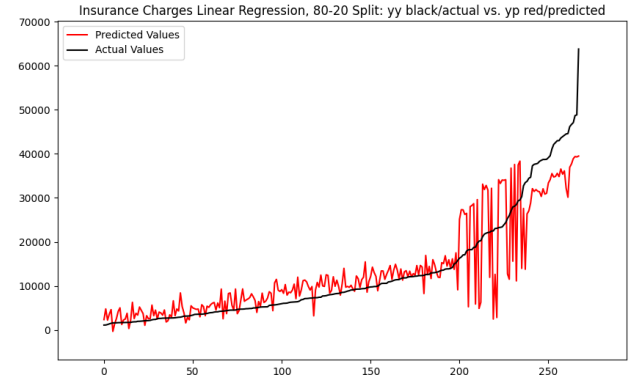
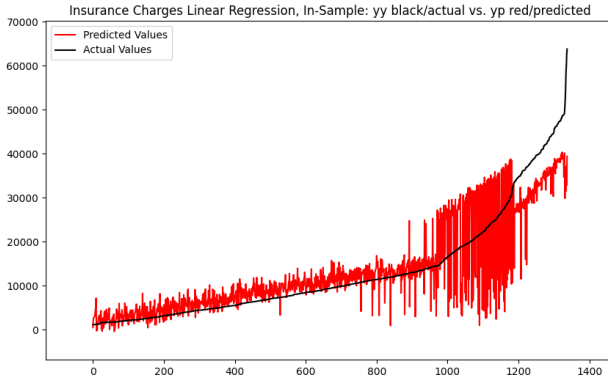


Figure 27: Statsmodels - Insurance Charges Lasso  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

## 5 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 lamda 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 from the modeling to perform 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)
```

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

Tables 13 and 14 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.

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

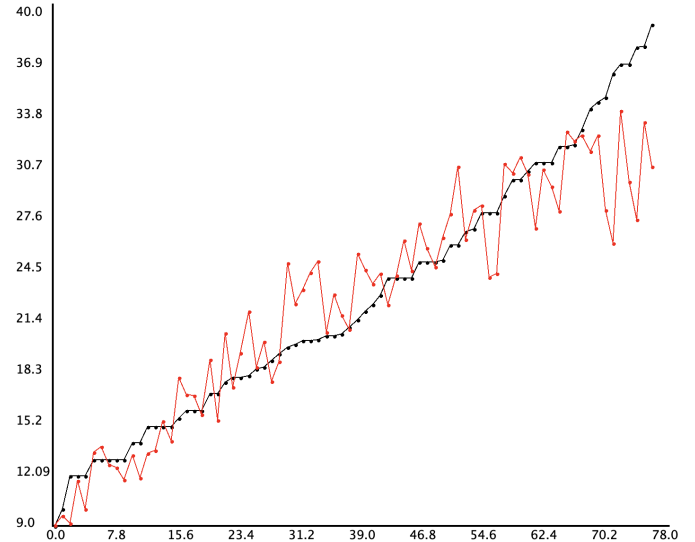
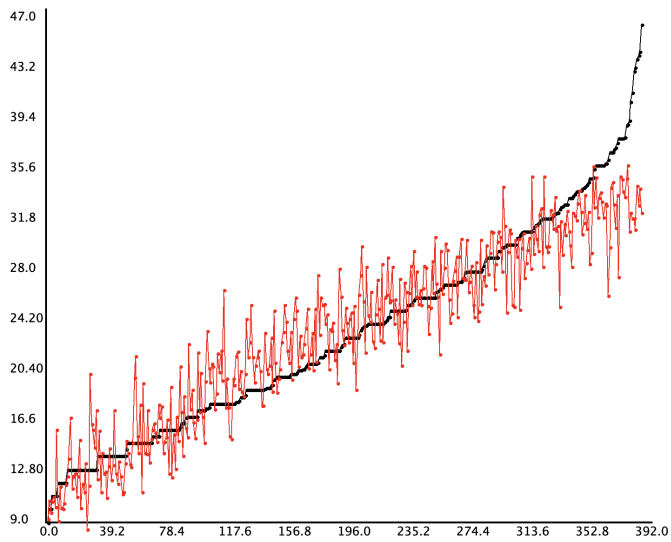


Figure 28: Scalation - Auto MPG Sqrt  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

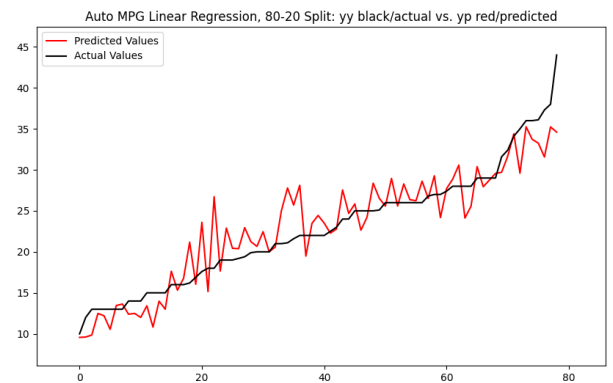
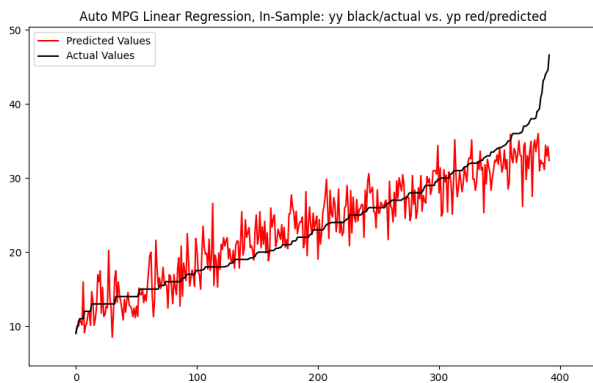


Figure 29: Statsmodels - Auto MPG Sqrt  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

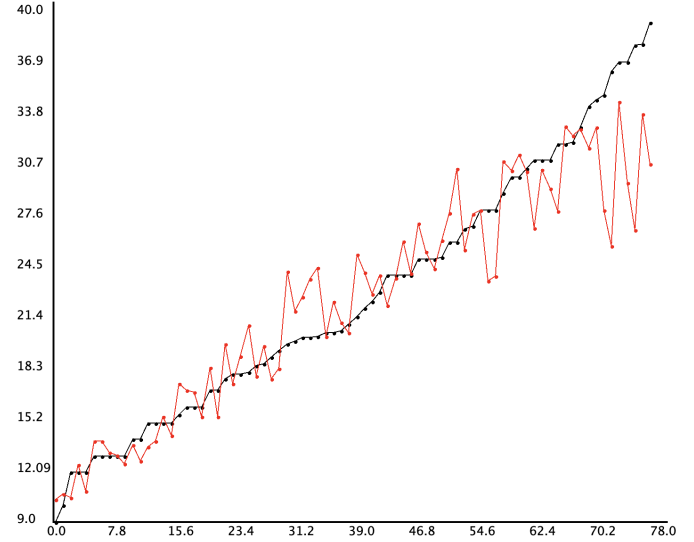
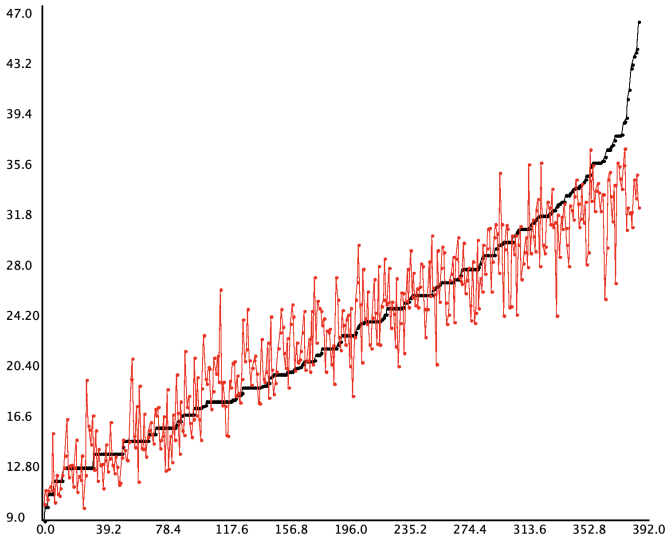


Figure 30: Scalation - Auto MPG log1p  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

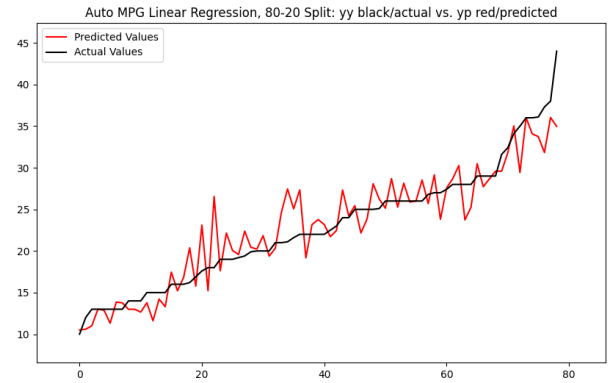
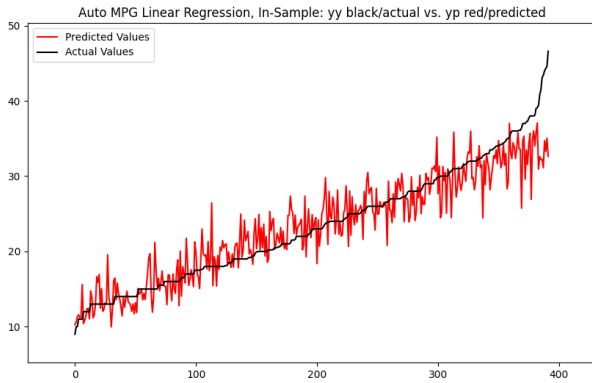


Figure 31: Statsmodels - Auto MPG log1p  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

## 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  $R^2$ .



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

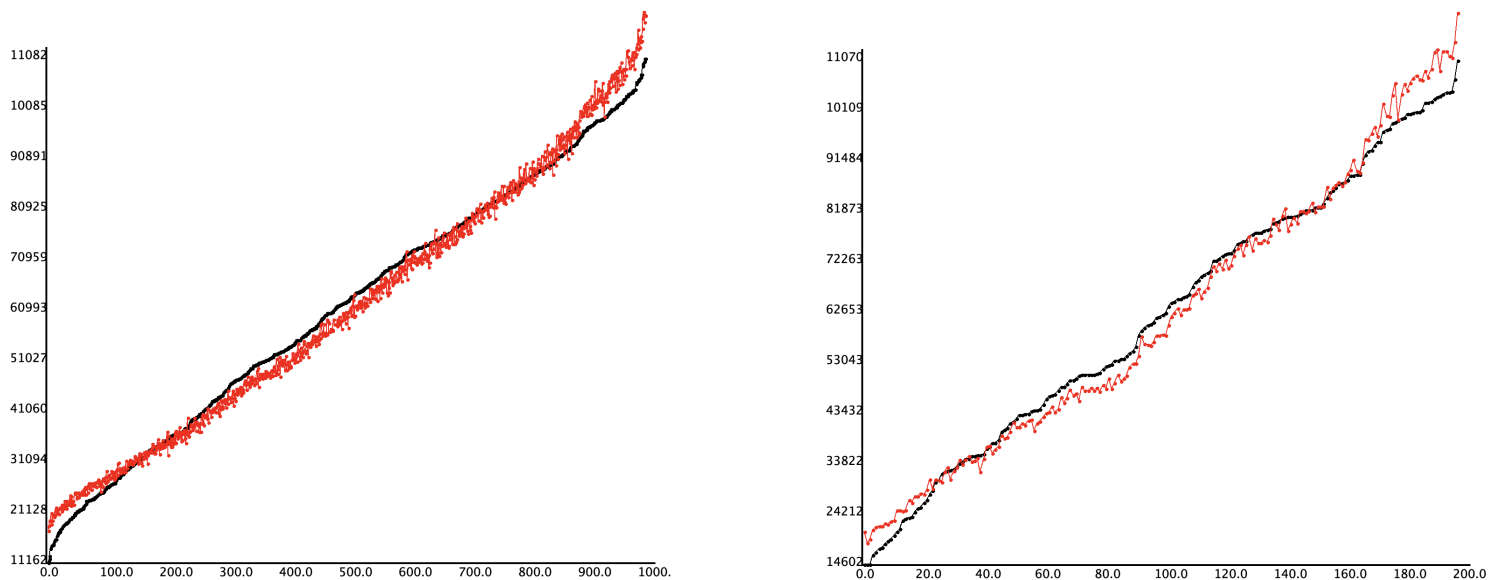


Figure 32: Scalation - House Price Sqrt  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

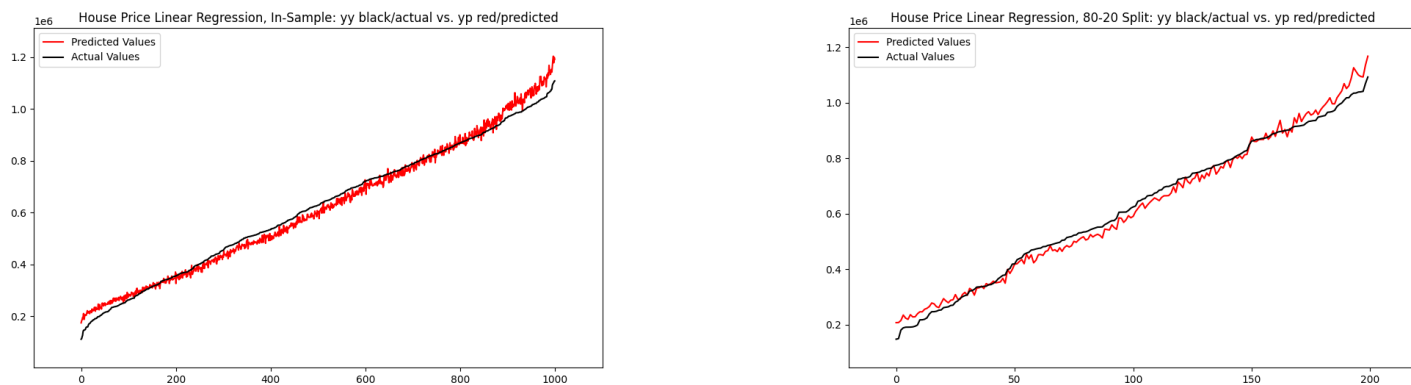


Figure 33: Statsmodels - House Price Sqrt  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

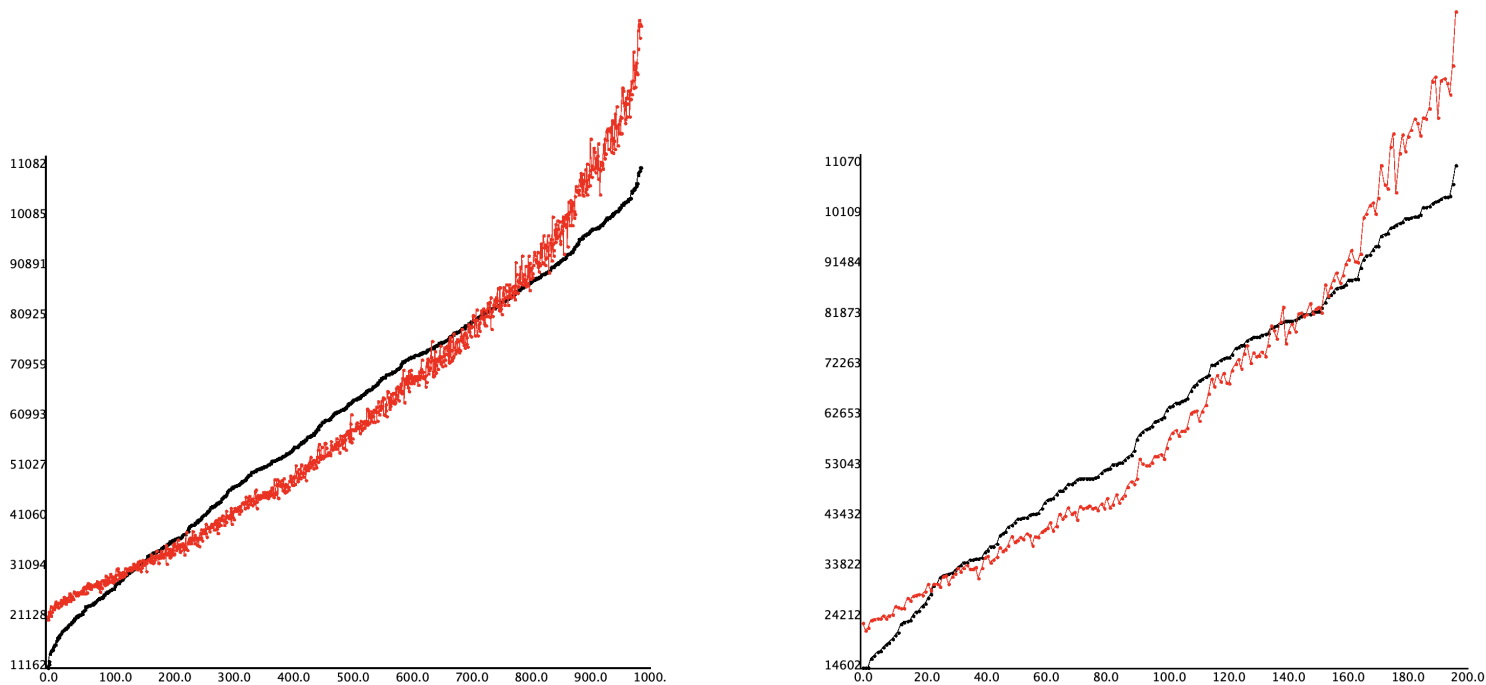


Figure 34: Scalation - House Price log1p  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

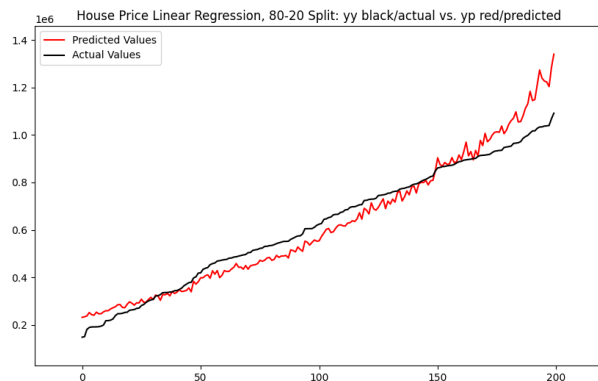
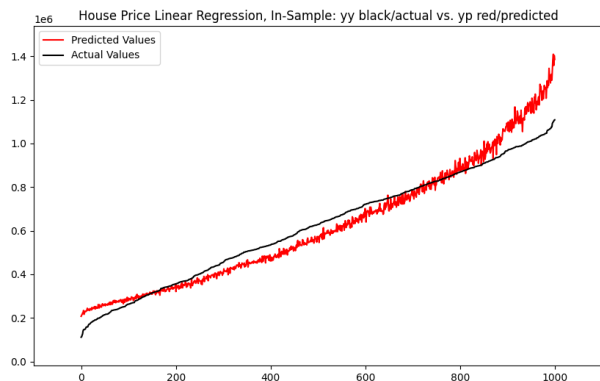


Figure 35: Statsmodels - House Price log1p  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

## 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.

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

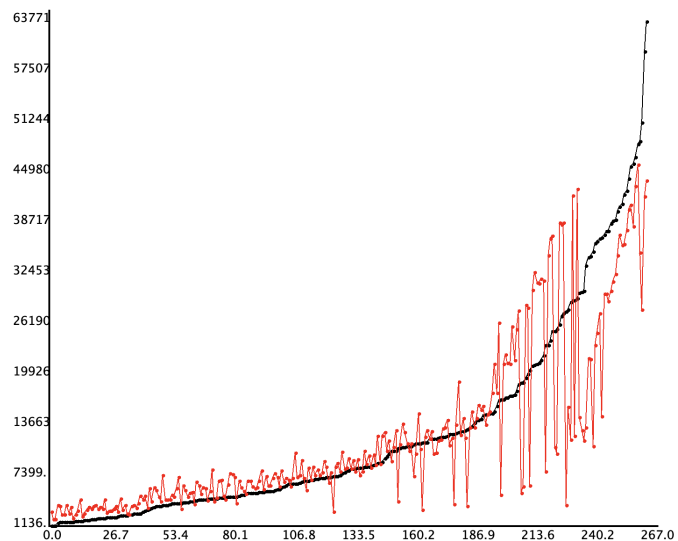
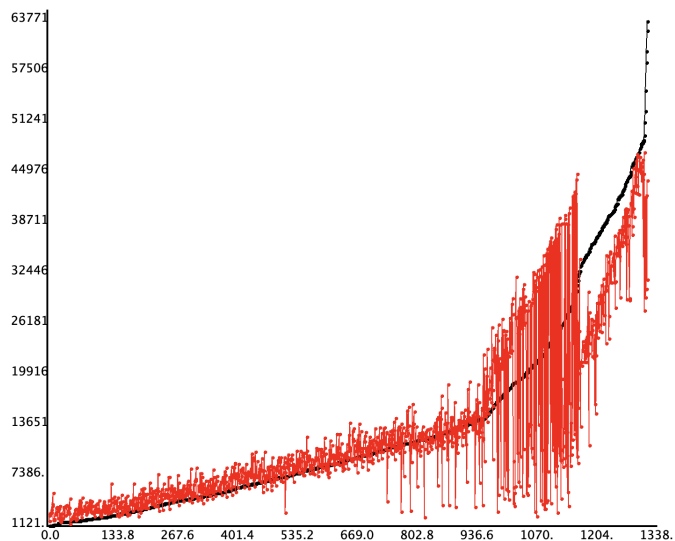


Figure 36: Scalation - Insurance Charges Sqrt  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

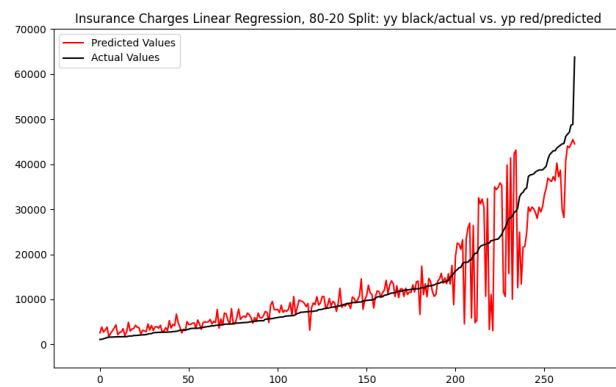
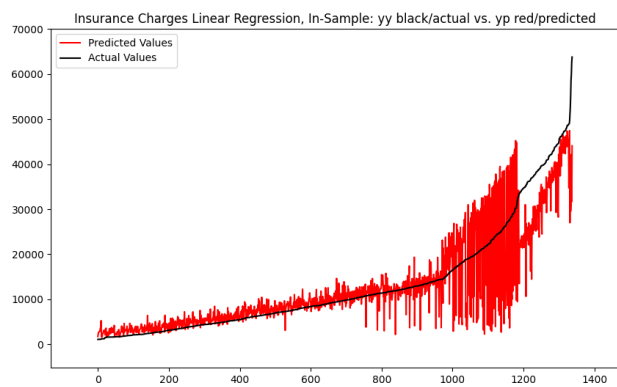


Figure 37: Statsmodels - Insurance Charges Sqrt  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

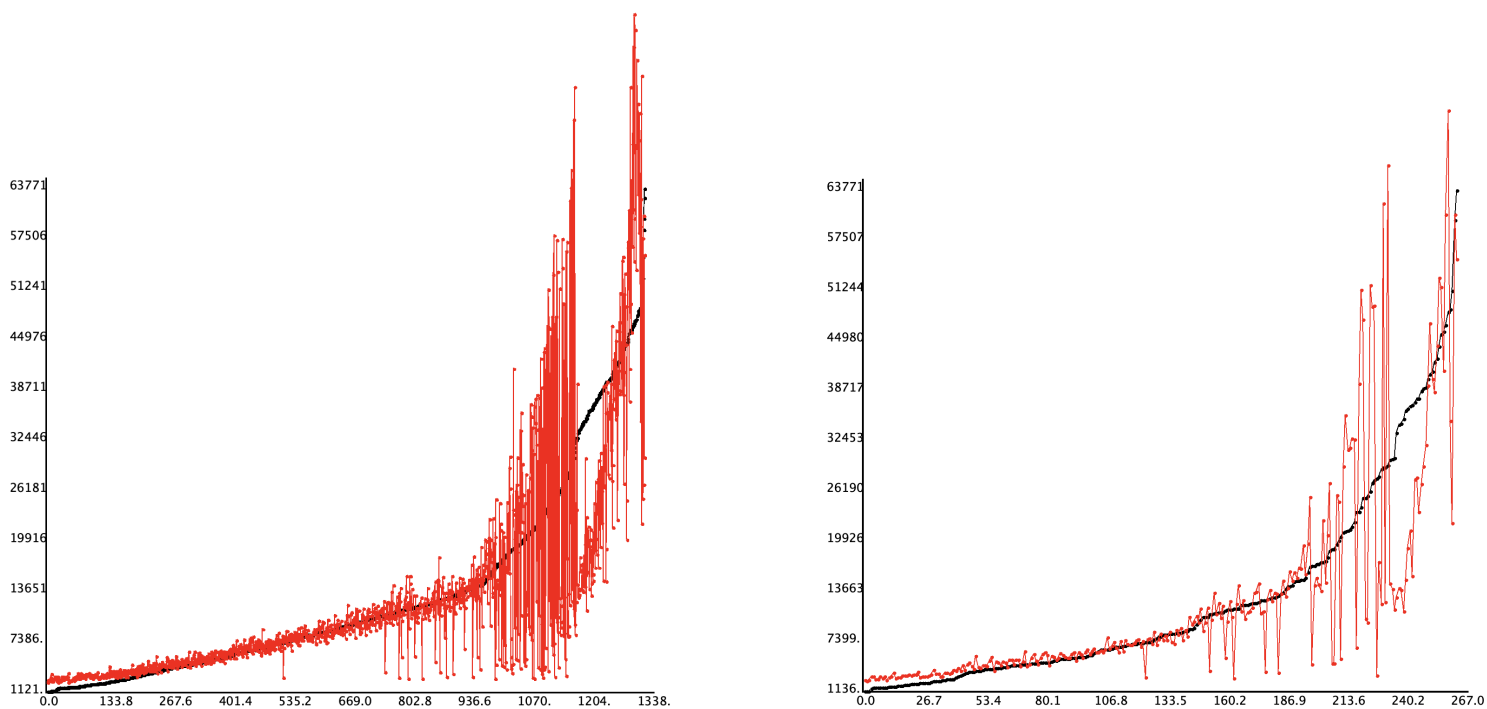


Figure 38: Scalation - Insurance Charges log1p  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

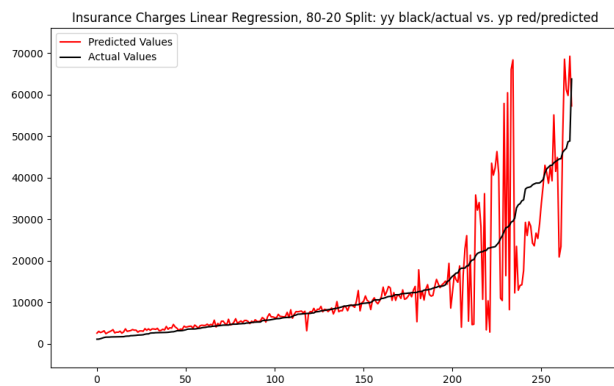
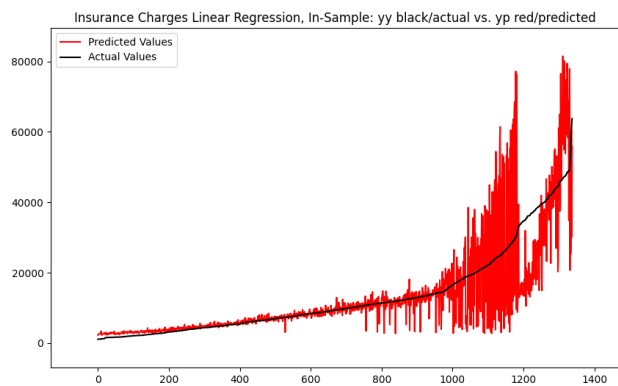


Figure 39: Statsmodels - Insurance Charges log1p  
 Left: In Sample Predictions  
 Right: 80-20 Out of Sample Predictions  
 yy black/actual vs. yp red/predicted

## 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.

Table 19: Auto MPG Regression with Square-Root Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.835	0.822	0.823	0.822
rSqBar	0.833	0.807	0.818	0.814
sst	23818.993	5372.801	5372.801	5372.801
sse	3926.849	955.390	950.390	958.806
mse	4.017	6.094	10.030	9.137
rmse	3.165	3.478	3.468	3.484
mae	2.347	2.376	2.375	2.350
m	392.000	79.000	79.000	79.000
dfr	6.000	6.000	2.000	3.000
df	385.000	72.000	76.000	75.000
fStat	825.235	120.822	220.454	161.034

Table 20: Auto MPG Regression with Log1p Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.849	0.832	0.806	0.804
rSqBar	0.847	0.818	0.801	0.794
sst	23818.993	5372.801	5372.801	5372.801
sse	3594.229	900.453	1044.122	1052.182
mse	3.169	5.398	11.217	9.319
rmse	3.028	3.376	3.635	3.649
mae	2.184	2.221	2.710	2.713
m	392.000	79.000	79.000	79.000
dfr	6.000	6.000	2.000	4.000
df	385.000	72.000	76.000	74.000
fStat	1063.694	138.083	192.956	115.912

Table 21: Auto MPG Regression with Box-Cox Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.845	0.830	0.829	0.829
rSqBar	0.843	0.816	0.825	0.823
sst	23818.993	5372.801	5372.801	5372.801
sse	3683.218	914.477	917.735	916.540
mse	3.396	5.576	9.617	8.602
rmse	3.065	3.402	3.408	3.406
mae	2.229	2.262	2.356	2.234
m	392.000	79.000	79.000	79.000
dfr	6.000	6.000	2.000	3.000
df	385.000	72.000	76.000	75.000
fStat	988.221	133.268	231.627	172.688

### 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.

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

	In-Sample	Validation	Forward	Backward
rSq	0.986	0.987	0.281	0.981
rSqBar	0.986	0.986	0.277	0.981
sst	64232463468052.547	11927062037792.469	11927062037792.469	11927062037792.469
sse	888710772890.930	155868297872.986	8579357295076.382	223122191452.058
mse	888710765.891	779341482.365	42896786474.382	1115610953.260
rmse	29811.252	27916.688	207115.394	33400.763
mae	24296.468	22632.355	177100.990	27022.336
m	1000.000	200.000	200.000	200.000
dfr	7.000	7.000	1.000	4.000
df	992.000	192.000	198.000	195.000
fStat	10182.286	2157.718	78.041	2622.765

Table 23: Boston House Price Regression with Log1p Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.922	0.934	-81398.531	0.877
rSqBar	0.922	0.931	-81809.639	0.875
sst	64232463468052.547	11927062037792.469	11927062037792.469	11927062037792.469
sse	4985983480190.929	789371506083.466	970857251871500928.000	1461933128041.826
mse	4985983473.191	3946857523.417	4854286259357504.000	7309665636.209
rmse	70611.497	62824.020	69672708.139	85496.583
mae	52989.293	48059.016	20766229.670	61057.028
m	1000.000	200.000	200.000	200.000
dfr	7.000	7.000	1.000	4.000
df	992.000	192.000	198.000	195.000
fStat	1697.515	403.131	-199.998	357.921



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

	In-Sample	Validation	Forward	Backward
rSq	0.998	0.997	0.924	0.990
rSqBar	0.998	0.997	0.924	0.989
sst	64232463468052.531	11927062037792.469	11927062037792.469	11927062037792.469
sse	157457720227.919	32215998366.736	903559871079.215	124824257960.295
mse	157457713.228	161079984.834	4517799354.396	624121285.801
rmse	12548.216	12691.729	67214.577	24982.420
mae	10078.898	10289.124	56408.842	20247.645
m	1000.000	200.000	200.000	200.000
dfr	7.000	7.000	1.000	4.000
df	992.000	192.000	198.000	195.000
fStat	58133.527	10549.192	2440.016	4727.542

### 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 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.

Table 25: Medical Cost Regression with Square-Root Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.753	0.706	0.702	0.704
rSqBar	0.751	0.697	0.699	0.699
sst	196074221568.367	35797001122.000	35797001122.000	35797001122.000
sse	48497528289.829	10509483431.295	10649651869.890	10582381967.672
mse	36246276.223	39214482.415	39737503.977	39486494.879
rmse	6020.489	6262.147	6303.769	6283.828
mae	3613.896	3611.073	3687.410	3606.738
m	1338.000	268.000	268.000	268.000
dfr	8.000	8.000	3.000	5.000
df	1329.000	259.000	264.000	262.000
fStat	508.937	80.606	210.946	127.713

Table 26: Medical Cost Regression with Log1p Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.523	0.505	-415.056	-53.303
rSqBar	0.520	0.490	-419.784	-54.340
sst	196074221568.367	35797001122.000	35797001122.000	35797001122.000
sse	93570168847.798	17725546380.471	14893568669580.479	1943894999500.171
mse	69932853.620	66140090.435	55573017420.808	7253339545.374
rmse	8362.587	8132.656	235739.300	85166.540
mae	4219.512	4231.425	66672.183	25327.152
m	1338.000	268.000	268.000	268.000
dfr	8.000	8.000	3.000	5.000
df	1329.000	259.000	264.000	262.000
fStat	183.219	34.154	-89.119	-52.613

Table 27: Medical Cost Regression with Box-Cox Transformation

	In-Sample	Validation	Forward	Backward
rSq	0.576	0.546	-61.221	-10.125
rSqBar	0.574	0.532	-61.928	-10.337
sst	196074221568.367	35797001122.000	35797001122.000	35797001122.000
sse	83083467025.754	16248889297.412	2227326392657.103	398225509782.019
mse	62095258.835	60630175.946	8310919372.586	1485916076.276
rmse	7880.055	7786.539	91164.244	38547.582
mae	4052.895	4083.531	31806.787	14527.915
m	1338.000	268.000	268.000	268.000
dfr	8.000	8.000	3.000	5.000
df	1329.000	259.000	264.000	262.000
fStat	227.454	40.302	-87.898	-48.782

## 6 Comparing Different Models on the Same Data Set

### 6.1 Auto MPG

Table 28: Scalation - Auto MPG In-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.809255	0.776580	0.809163	0.835138	0.849102
rSqBar	0.806283	0.772507	0.806189	0.832569	0.846751
sst	23819.0	23819.0	23819.0	23819.0	23819.0
sse	4543.35	5321.63	4545.54	3926.85	3594.23
sde	3.40878	3.68883	3.40888	3.16757	3.02514
mse0	11.5902	13.5756	11.5958	10.0175	9.16895
rmse	3.40443	3.68451	3.40526	3.16504	3.02803
mae	2.61826	2.79509	2.61703	2.34675	2.18422
smape	12.0589	65.4181	11.9861	10.2046	9.32001
m	392.000	392.000	392.000	392.000	392.000
dfr	6.00000	7.00000	6.00000	6.00000	6.00000
df	385.000	384.000	385.000	385.000	385.000
fStat	272.234	190.677	272.072	325.048	361.067
aic	-1022.45	-1051.45	-1022.55	-993.873	-976.527
bic	-994.656	-1019.68	-994.750	-966.074	-948.728

Table 29: Scalation - Auto MPG Out-of-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.822842	0.797491	0.822903	0.846480	0.852864
rSqBar	0.820081	0.793799	0.820143	0.844088	0.850571
sst	4731.23	4731.23	4731.23	4731.23	4731.23
sse	838.174	958.118	837.889	726.337	696.133
sde	3.29026	3.51709	3.28969	3.05724	2.97969
mse0	10.7458	12.2836	10.7422	9.31202	8.92478
rmse	3.27808	3.50479	3.27752	3.05156	2.98744
mae	2.48735	2.62052	2.48643	2.11846	1.98665
smape	11.8858	61.9272	11.8808	9.00068	8.28831
m	78.0000	78.0000	78.0000	78.0000	78.0000
dfr	6.00000	7.00000	6.00000	6.00000	6.00000
df	385.000	384.000	385.000	385.000	385.000
fStat	298.034	216.030	298.158	353.804	371.939
aic	-189.284	-192.500	-189.271	-183.700	-182.048
bic	-172.787	-173.646	-172.774	-167.203	-165.551

Table 30: Statsmodels - Auto MPG In-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	Log1p
rSq	0.8093	0.8037	0.6416	0.8477	0.8725
rSqBar	0.8063	0.8012	0.6369	0.8453	0.8705
sst	23818.9935	23818.9935	23818.9935	252.1610	41.1728
sse	4543.3470	4675.2421	8537.0473	38.4071	5.2483
sde	3.4352	3.4802	4.7028	0.3158	0.1168
mse0	11.8009	11.9266	21.7782	0.0998	0.0136
rmse	3.4352	3.4535	4.6667	0.3158	0.1168
mae	2.6183	2.6269	3.6246	2.3467	2.1842
smape	12.0589	12.0433	16.2905	10.2046	9.3200
m	392.0000	392.0000	392.0000	392.0000	392.0000
dfr	6.0000	6.0000	6.0000	6.0000	6.0000
df	385.0000	386.0000	386.0000	385.0000	385.0000
fStat	272.2341	263.4262	115.1614	357.1178	439.2179
aic	2086.9095	983.6796	1219.7162	215.8245	-564.3874
bic	2114.7083	1007.5071	1243.5438	243.6234	-536.5886

Table 31: Statsmodels - Auto MPG Out-of-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	Log1p
rSq	0.8107	0.7854	0.5925	0.8482	0.8720
rSqBar	0.8070	0.7707	0.5646	0.8452	0.8695
sst	19750.2199	4032.2061	4032.2061	208.0819	33.8141
sse	3738.2664	865.2680	1643.1024	31.5951	4.3287
sde	3.4952	3.4428	4.7443	0.3213	0.1189
mse0	12.2166	10.9528	20.7988	0.1033	0.0141
rmse	3.4952	3.3095	4.5606	0.3213	0.1189
mae	2.5039	2.5877	3.7495	2.1509	1.9533
smape	12.3880	12.5974	17.6830	9.8913	8.7248
m	79.0000	79.0000	79.0000	79.0000	79.0000
dfr	6.0000	6.0000	6.0000	6.0000	6.0000
df	306.0000	73.0000	73.0000	306.0000	306.0000
fStat	218.4461	44.5308	17.6906	284.8805	347.3956
aic	1678.5500	201.0937	251.7566	184.4835	-437.6799
bic	1704.7734	215.3104	265.9733	210.7069	-411.4565

## 6.2 House Prices

Table 32: Scalation - Housing Prices In-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.998516	0.987398	0.998516	0.986152	0.922307
rSqBar	0.998507	0.987309	0.998507	0.986068	0.921838
sst	6.42325e+13	6.42325e+13	6.42325e+13	6.42325e+13	6.42325e+13
sse	9.53030e+10	8.09438e+11	9.53030e+10	8.89504e+11	4.99039e+12
sde	9767.21	28463.5	9767.21	29836.4	70604.5
mse0	9.53030e+07	8.09438e+08	9.53030e+07	8.89504e+08	4.99039e+09
rmse	9762.32	28450.6	9762.32	29824.5	70642.7
mae	7747.66	24147.7	7747.66	24309.3	53070.5
smape	1.57791	23.5576	1.57791	4.86164	9.22319
m	1000.00	1000.00	1000.00	1000.00	1000.00
dfr	6.00000	7.00000	6.00000	6.00000	6.00000
df	993.000	992.000	993.000	993.000	993.000
fStat	111378	11103.9	111378	11785.5	1964.69
aic	-10591.2	-11658.9	-10591.2	-11708.0	-12570.3
bic	-10556.9	-11619.6	-10556.9	-11673.7	-12536.0

Table 33: Scalation - Housing Prices Out-of-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.998649	0.989193	0.998649	0.984018	0.906860
rSqBar	0.998641	0.989117	0.998641	0.983921	0.906297
sst	1.33700e+13	1.33700e+13	1.33700e+13	1.33700e+13	1.33700e+13
sse	1.80638e+10	1.44485e+11	1.80638e+10	2.13678e+11	1.24528e+12
sde	9501.56	26880.4	9501.56	32757.9	78673.0
mse0	9.03190e+07	7.22424e+08	9.03189e+07	1.06839e+09	6.22641e+09
rmse	9503.63	26877.9	9503.63	32686.3	78907.6
mae	7484.48	22419.4	7484.48	26186.8	58148.8
smape	1.60847	19.6544	1.60847	5.19788	9.91629
m	200.000	200.000	200.000	200.000	200.000
dfr	6.00000	7.00000	6.00000	6.00000	6.00000
df	993.000	992.000	993.000	993.000	993.000
fStat	122330	12971.9	122330	10189.9	1611.39
aic	-2101.67	-2307.60	-2101.67	-2348.73	-2525.00
bic	-2078.59	-2281.21	-2078.59	-2325.64	-2501.91

Table 34: Statsmodels - House Price In-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	Log1p
rSq	0.9985	0.9906	0.9875	0.9855	0.9415
rSqBar	0.9985	0.9905	0.9875	0.9854	0.9411
sst	64232463468052.5469	64232463468052.5469	64232463468052.5469	29445342.0661	241.7492
sse	95249090298.3967	604344645977.1542	800169699033.4265	425598.9981	14.1461
sde	9798.8381	24669.9185	28386.7992	20.7131	0.1194
mse0	96017228.1234	604344645.9772	800169699.0334	429.0312	0.0143
rmse	9798.8381	24583.4222	28287.2710	20.7131	0.1194
mae	7740.4301	20205.2499	24002.5173	24296.4685	52989.2927
smape	1.5774	4.1035	4.9163	4.8579	9.2122
m	1000.0000	1000.0000	1000.0000	1000.0000	1000.0000
dfr	7.0000	7.0000	7.0000	7.0000	7.0000
df	992.0000	993.0000	993.0000	992.0000	992.0000
fStat	95425.1583	14935.3572	11245.5195	9662.8803	2280.1137
aic	21225.8831	20233.6552	20514.3344	8907.3747	-1404.4424
bic	21265.1451	20268.0095	20548.6887	8946.6367	-1365.1804

Table 35: Statsmodels - House Price Out-of-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	Log1p
rSq	0.9985	0.9913	0.9884	0.9852	0.9402
rSqBar	0.9985	0.9910	0.9880	0.9851	0.9397
sst	51340367232368.6250	12891771417242.6445	12891771417242.6445	23544096.3448	193.6629
sse	75080425564.8099	112341107192.6465	149469076702.9745	347277.4580	11.5792
sde	9736.4530	24126.2984	27828.9629	20.9400	0.1209
mse0	94798517.1273	561705535.9632	747345383.5149	438.4816	0.0146
rmse	9736.4530	23700.3278	27337.6185	20.9400	0.1209
mae	8174.5836	19629.7981	23012.0161	23553.0274	51521.6937
smape	1.6620	3.9691	4.6112	4.7594	9.0567
m	200.0000	200.0000	200.0000	200.0000	200.0000
dfr	7.0000	7.0000	7.0000	7.0000	7.0000
df	792.0000	193.0000	193.0000	792.0000	792.0000
fStat	77254.5038	3136.4045	2350.4760	7557.5143	1779.1786
aic	16972.0728	4043.2977	4100.4076	7144.9157	-1102.0191
bic	17009.5497	4066.3859	4123.4958	7182.3926	-1064.5422

### 6.3 Insurance Charges

Table 36: Scalation - Insurance Charges In-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.750157	0.749962	0.750157	0.752656	0.527431
rSqBar	0.748842	0.748457	0.748842	0.751354	0.524943
sst	1.96074e+11	1.96074e+11	1.96074e+11	1.96074e+11	1.96074e+11
sse	4.89878e+10	4.90260e+10	4.89878e+10	4.84977e+10	9.26587e+10
sde	6053.11	6055.42	6053.11	6001.45	8316.89
mse0	3.66127e+07	3.66413e+07	3.66127e+07	3.62464e+07	6.92516e+07
rmse	6050.84	6053.20	6050.84	6020.50	8321.76
mae	4179.54	4150.46	4179.54	3623.20	4215.04
smape	37.9722	67.8472	37.9722	27.9341	26.5034
m	1338.00	1338.00	1338.00	1338.00	1338.00
dfr	7.00000	8.00000	7.00000	7.00000	7.00000
df	1330.00	1329.00	1330.00	1330.00	1330.00
fStat	570.477	498.274	570.477	578.161	212.057
aic	-13533.8	-13532.3	-13533.8	-13527.1	-13960.2
bic	-13492.2	-13485.5	-13492.2	-13485.5	-13918.6

Table 37: Scalation - Insurance Charges Out-of-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	log1p
rSq	0.720005	0.719714	0.720005	0.730926	0.583107
rSqBar	0.718531	0.718027	0.718531	0.729510	0.580913
sst	4.06432e+10	4.06432e+10	4.06432e+10	4.06432e+10	4.06432e+10
sse	1.13799e+10	1.13917e+10	1.13799e+10	1.09360e+10	1.69438e+10
sde	6540.46	6543.82	6540.46	6396.98	7972.86
mse0	4.26213e+07	4.26655e+07	4.26213e+07	4.09588e+07	6.34601e+07
rmse	6528.50	6531.89	6528.50	6399.91	7966.18
mae	4430.66	4429.38	4430.66	3868.65	4087.97
smape	40.0602	62.9178	40.0602	30.5895	27.5909
m	267.000	267.000	267.000	267.000	267.000
dfr	7.00000	8.00000	7.00000	7.00000	7.00000
df	1330.00	1329.00	1330.00	1330.00	1330.00
fStat	488.584	426.574	488.584	516.126	265.753
aic	-2708.17	-2706.31	-2708.17	-2702.86	-2761.31
bic	-2679.47	-2674.02	-2679.47	-2674.16	-2732.61

Table 38: Statsmodels - Insurance Charges In-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	Log1p
rSq	0.7509	0.2718	0.7469	0.7795	0.7680
rSqBar	0.7494	0.2680	0.7456	0.7782	0.7666
sst	196074221568.3671	196074221568.3671	196074221568.3671	3051091.5131	1130.1100
sse	48839532843.9219	142780871945.3538	49623455821.6671	672635.9707	262.2315
sde	6062.1023	10361.1794	6108.2624	22.4972	0.4442
mse0	36749084.1564	106712161.3941	37087784.6201	506.1219	0.1973
rmse	6062.1023	10330.1579	6089.9741	22.4972	0.4442
mae	4170.8869	8134.7714	4115.6884	3613.8958	4219.5115
smape	37.8059	66.1576	34.1997	27.6903	26.2889
m	1338.0000	1338.0000	1338.0000	1338.0000	1338.0000
dfr	8.0000	8.0000	8.0000	8.0000	8.0000
df	1329.0000	1330.0000	1330.0000	1329.0000	1329.0000
fStat	500.8107	62.0533	490.6438	587.4216	549.8054
aic	27113.5058	24749.7939	23335.7320	12137.4774	1634.5362
bic	27160.2962	24791.3854	23377.3235	12184.2678	1681.3266

Table 39: Statsmodels - Insurance Charges Out-of-Sample QoF Comparison

Metric	Regression	Ridge	Lasso	Sqrt	Loglp
rSq	0.7417	0.2951	0.7797	0.7701	0.7572
rSqBar	0.7398	0.2761	0.7738	0.7684	0.7554
sst	154436975468.4681	41606660039.7953	41606660039.7953	2399231.6774	888.0691
sse	39887119421.1600	29328394351.2134	9164643207.2381	551569.9183	215.5882
sde	6131.3858	10620.8058	5937.0555	22.8004	0.4508
mse0	37593892.0086	109434307.2806	34196429.8778	519.8585	0.2032
rmse	6131.3858	10461.0854	5847.7714	22.8004	0.4508
mae	4181.1945	8328.6585	4069.5309	3556.9640	3888.4432
smape	40.0220	69.4557	35.5690	29.1245	25.7136
m	268.0000	268.0000	268.0000	268.0000	268.0000
dfr	8.0000	8.0000	8.0000	8.0000	8.0000
df	1061.0000	260.0000	260.0000	1061.0000	1061.0000
fStat	380.8792	13.6061	115.0471	444.2703	413.6950
aic	21708.8072	4976.9038	4665.1653	9736.7961	1340.3418
bic	21753.5859	5005.6317	4693.8932	9781.5748	1385.1205